PROGRAMME AND ABSTRACTS

26th International Conference on Computational Statistics (COMPSTAT 2024)

http://www.compstat2024.org

Justus-Liebig-University of Giessen, Germany
27-30 August 2024

ISBN: 9789073592421

Financial support of the conference by DFG through project number 537703842 is gratefully acknowledged.

COMPSTAT 2024 Scientific Program Committee:

Ex-officio:

COMPSTAT 2024 organiser and chairperson of the SPC: Peter Winker and Ana Colubi.
Past COMPSTAT organiser: Erricos Kontoghiorghes.
Next COMPSTAT organiser: Ioannis Demetriou.
Incoming IASC-ERS Chairman: Ana Belen Ramos-Guajardo.

Members:

Alessandra Amendola, Stefanie Biedermann, Marc Hallin, Stefan Sperlich and Mattias Villani.

Consultative Members:

Representative of the IFCS: Rebecca Nugent.
Representative of the ARS of IASC: Ray-Bing Chen.
Representative of the LARS of IASC: Veronica A. Gonzalez-Lopez..
Representative of CMStatistics: Erricos Kontoghiorghes.

Local Organizing Committee:

Roland Fried, Sonja Greven, Roxana Halbleib, Claudia Kirch, Thomas Kneib and Dominik Liebl.

Dear Colleagues and Friends,

Welcome to the 26th International Conference on Computational Statistics (COMPSTAT 2024) in Giessen. This remarkable edition of COMPSTAT celebrates its 50th anniversary. COMPSTAT 2024 is set to commemorate this milestone during the opening ceremony. Lutz Edler will provide a brief overview of the conference's history during the opening ceremony, reflecting on its journey and contributions over the past five decades.

The organization of this event has been primarily led by members of Justus-Liebig-University of Giessen, 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 35 contributed sessions, 6 invited sessions, 3 keynote talks, 37 organized sessions, and one extended tutorial, with approximately 360 participants. To accommodate various preferences, COMPSTAT 2024 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 2024. 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 Athens, 25-28 August 2026, for the 27th edition of COMPSTAT. Our best wishes for success go to the chair of the upcoming edition.

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

Peter Winker and Ana Colubi
Organisers and chairpersons

SCHEDULE COMPSTAT 2024

2024-08-27

Opening
09:10 - 09:40

A - Keynote
COMPSTAT2024
09:40 - 10:30

Coffee break
10:30 - 11:00

B
COMPSTAT2024
11:00 - 12:30

Lunch break
12:30 - 14:00

C
COMPSTAT2024
14:00 - 15:30

Coffee break
15:30 - 16:00

D
COMPSTAT2024
16:00 - 18:00

Welcome reception
18:00 - 19:30

2024-08-28

E
COMPSTAT2024
09:00 - 10:30

Coffee break
10:30 - 11:00

F
COMPSTAT2024
11:00 - 12:30

Lunch break
12:30 - 14:00

G
COMPSTAT2024
14:00 - 16:00

Guided visit
16:30 - 18:30

2024-08-29

H
COMPSTAT2024
09:00 - 10:00

Coffee break
10:00 - 10:30

I
COMPSTAT2024
10:30 - 12:30

Lunch break
12:30 - 14:00

J
COMPSTAT2024
14:00 - 15:30

Coffee break
15:30 - 16:00

K
COMPSTAT2024
16:00 - 17:30

L - Keynote
COMPSTAT2024
17:40 - 18:30

Conference dinner
19:30 - 22:00

2024-08-30

M
COMPSTAT2024
09:00 - 10:30

Coffee break
10:30 - 11:00

N
COMPSTAT2024
11:00 - 12:00

O - Keynote
COMPSTAT2024
12:10 - 13:00

Lunch break
13:00 - 14:00

P
COMPSTAT2024
14:00 - 16:30

General information, tutorial, and social events

Address of venues

The conference venue is the Lecture Hall Building, Law and Economics Building (Hörsaalgebäude Recht und Wirtschaft), Justus-Liebig-University of Giessen, Licher Strasse 68, 35394 Giessen.

Registration

The registration will be open on Monday from 17:00 to 19:00, during the ice-breaker, and each day, from Tuesday, the 27th of August 2024, from 08:30 until the end of the sessions. It will take place in the hall to the right of Auditorium HS 4, on the ground floor of the old part of the building.

Presentation instructions

The keynote talks will take place in the Auditorium 4, on the ground floor of the old part of the building. The rest of the conference will run in the new part of the building (ground floor, semi-basement -1, and semi-basement -2). The poster sessions will take place online, but in-person participants can meet in the designated room. 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 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

Posters will be displayed on Zoom. Presenters should select the breakout room with their poster code (e.g. E0123), share the poster and remain with the camera, microphone, chat and audio ready throughout the entire session. 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, the 24th of August 2024, from 15:00 to 15:30. Participants will be able to enter through the Room 050 to test their presentations, video, microphone, and audio (e.g., through Parallel session B). Detailed indications for the test sessions can be found on the website.

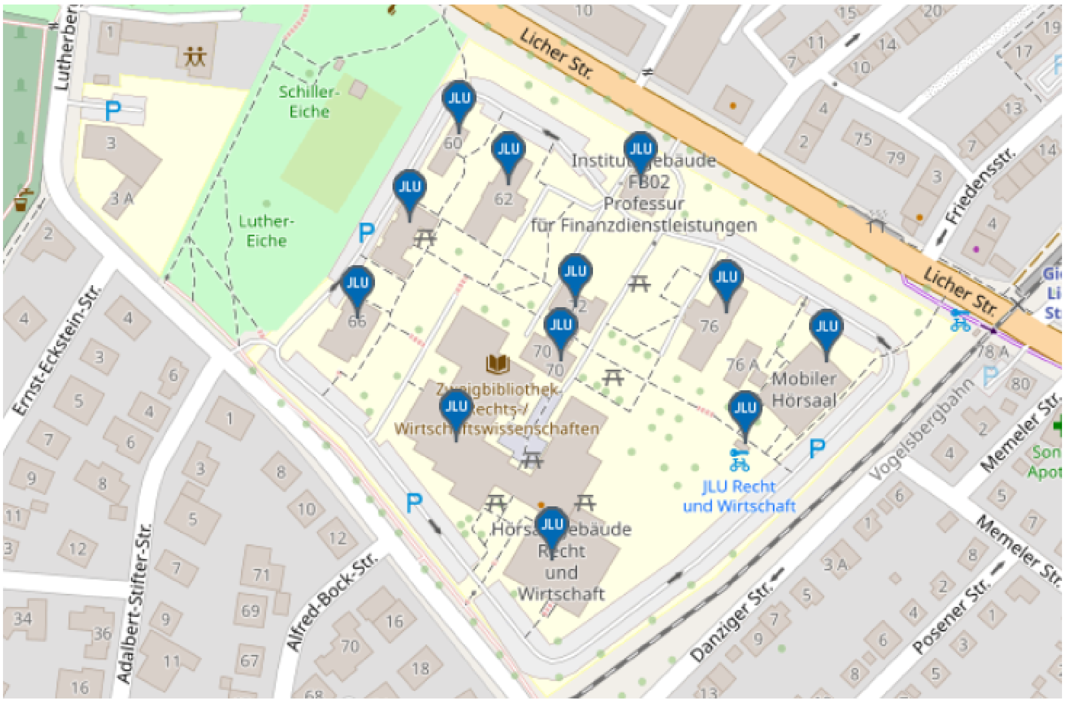
HiTEc and COMPSTAT 2024 tutorial

The 5-hour tutorial “Topic modelling” will take place on the 30th of August in Room 44. It has been organized within COMPSTAT by the COST Action HiTEc and it will be delivered by Dr. Ivan Savin. Details can be found on the website.

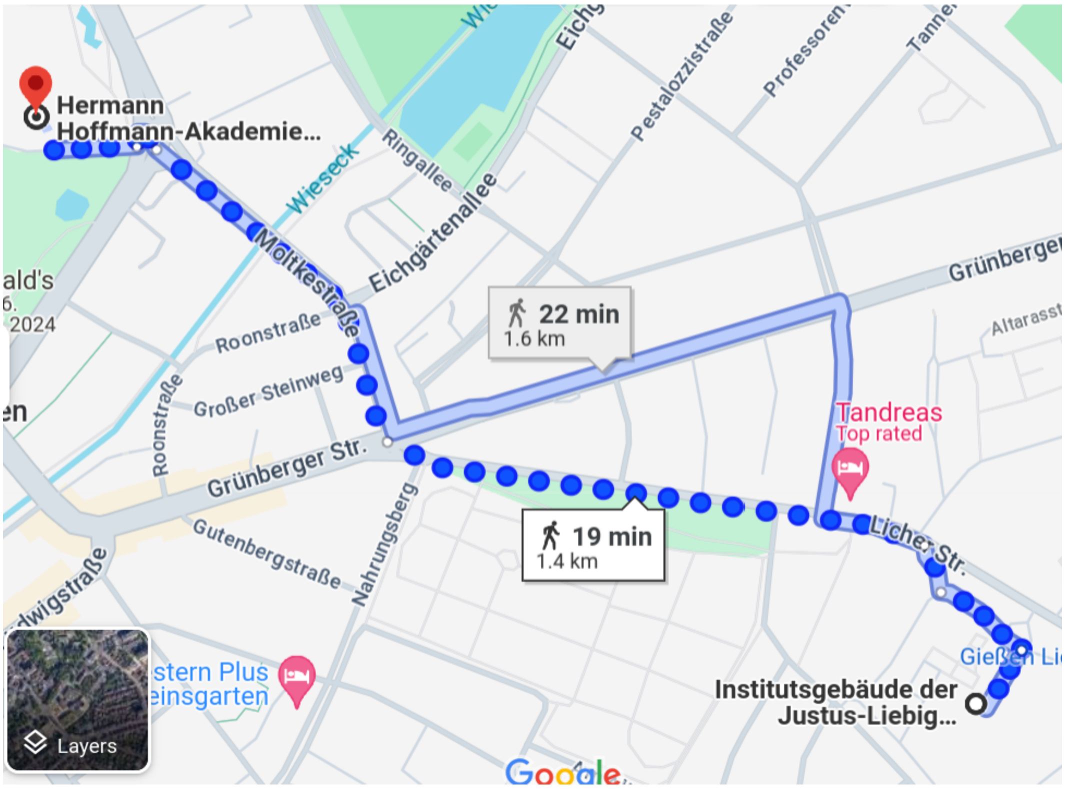
SOCIAL EVENTS

- *Ice-breaker, Monday, the 26th of August 2024, 17:00-19:00:* Hall on the right of Auditorium HS 4, on the ground floor of the old part of the building. It is open to all participants, who will be able to register at the same time.
- *The coffee breaks:* Hall on the right of Auditorium HS 4, on the ground floor of the old part of the building. Participants must have their conference badge in order to attend the coffee breaks.
- *Lunches:* University canteen, for those who had booked. Lunches include a main dish (vegan and vegetarian options available), a side dish, a dessert and a soft drink. They are optional and registration is required. You must book the corresponding lunch when registering and bring your badge in order to access the lunches each day. People not registered for lunch can buy lunch at restaurants and cafes within walking distance to the conference venue.
- *Welcome Reception, Tuesday the 27th of August 2024, 18:00-19:30.* Hall to the right of Auditorium HS 4, on the ground floor of the old part of the building. Open to all registrants who had preregistered and accompanying persons who have purchased a reception ticket. Preregistration is required due to health and safety reasons. Participants must bring their conference badge in order to attend the reception.
- *Guided visit, 28th of August 2024, 16:30-18:30:* Senkenbergstrasse 1721. The meeting point is at the entrance shown with the dinosaur. Those participants who are in the venue at 16:00 can meet by the registration desk. The conference staff will guide them to the meeting point. The visit is optional, and registration is required. Participants must bring their conference badge in order to attend the event.
- *Conference Dinner, Thursday, 29th of August 2024, 19:30-22:00:* Hotel & Restaurant heyligenstaedt, Aulweg 41, 35392 GieSSen. The conference dinner is optional, and registration is required. Participants must bring their conference badges to attend the conference dinner.

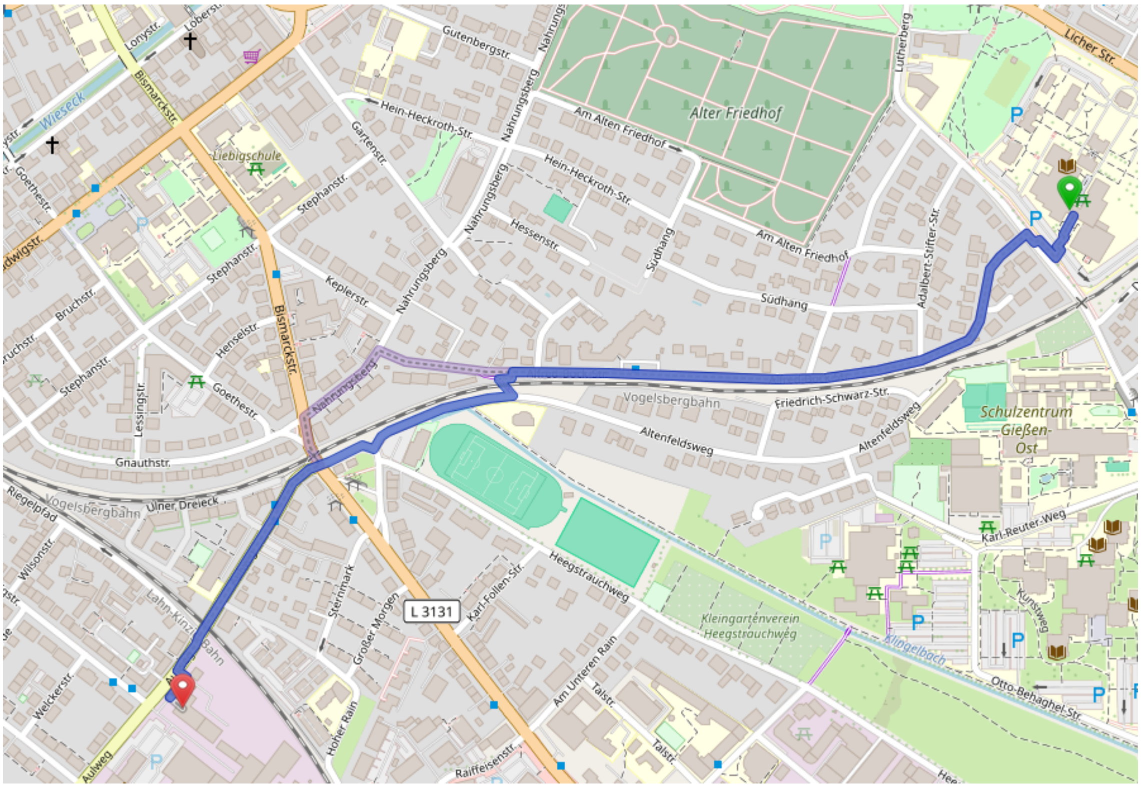
COMPSTAT venue



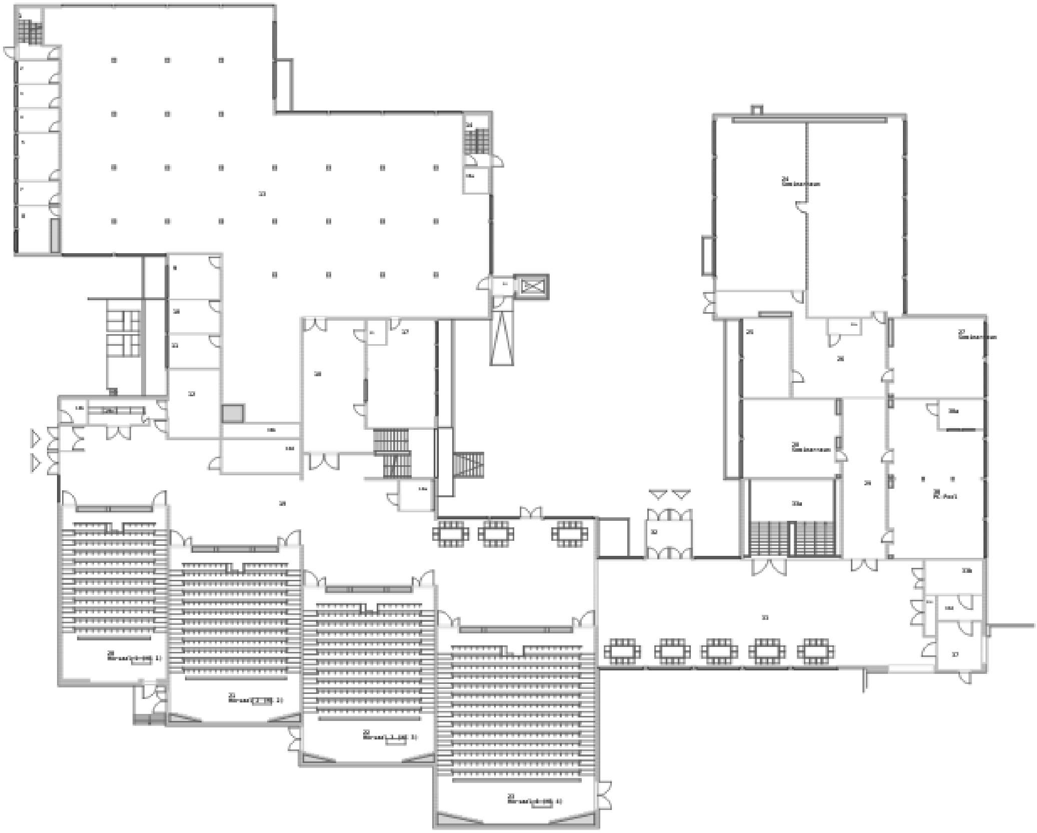
Guided visit



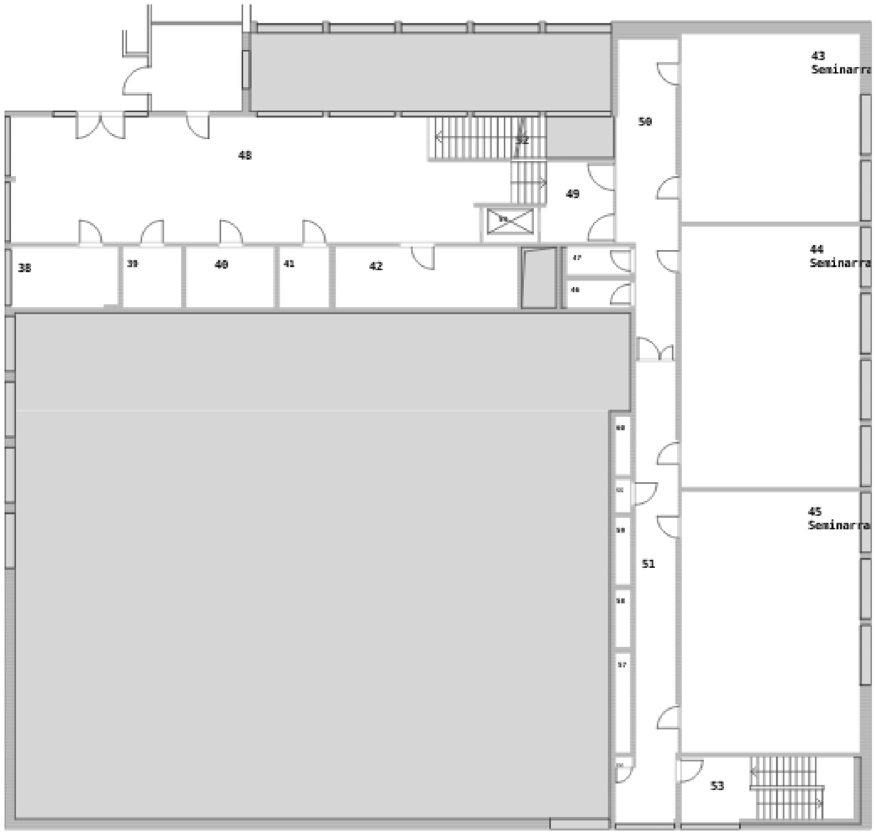
Conference dinner



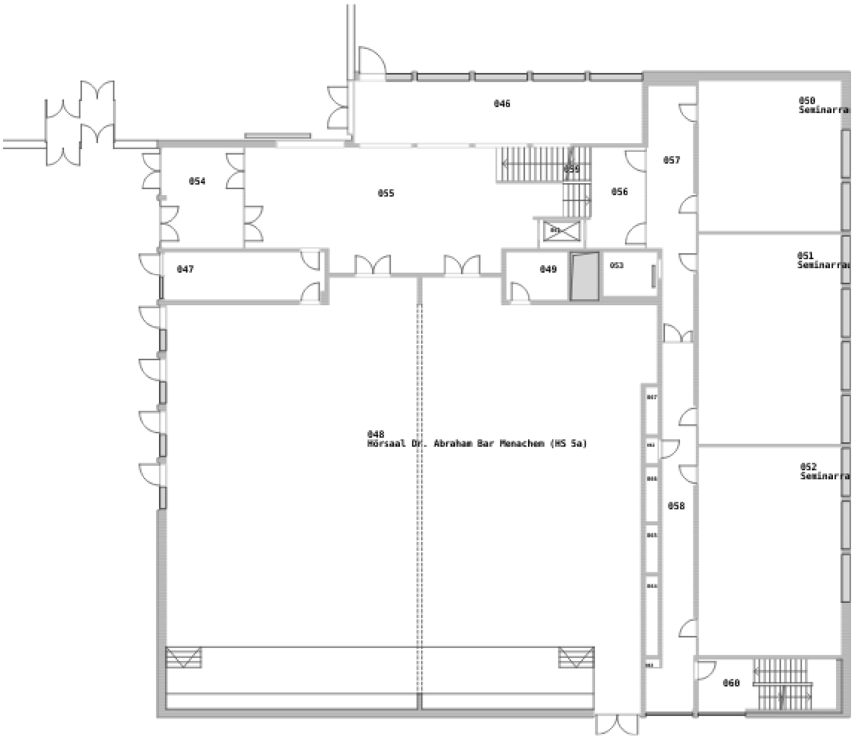
Ground floor (EG old building)



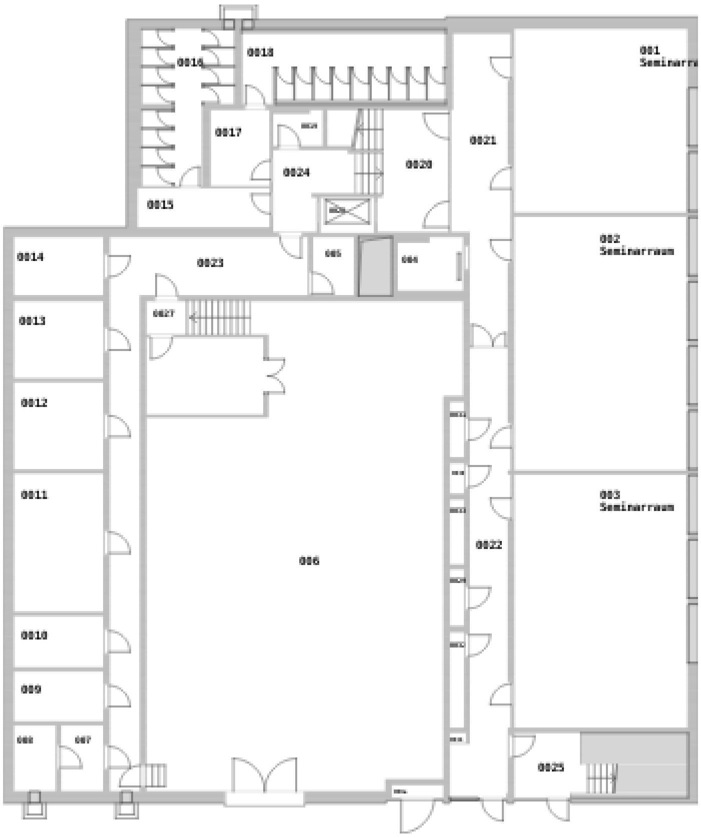
Ground floor (EG new building)



Semi-basement -1 (EG-1)



Semi-basement -2 (EG -2)



Contents

Tuesday 27.08.2024 09:40 - 10:30 Room: Auditorioum 4 Chair: Peter Winker Keynote talk I

On regression with convex classes

Speaker: **Sara van de Geer, ETH Zurich, Switzerland**

Least squares estimation is considered over a convex class of regression functions that can be well-approximated by linear functions. We assume that the dimension $M(\varepsilon)$ needed for a linear ε-approximation of the class grows polynomially in $1/\varepsilon$, with exponent $W > 0$. In that case, the rate of convergence of the least squares estimator is up to log-terms of order $n^{-\frac{2+W}{2(1+W)}}$ where n is the number of observations. The result is applied to the case where the class of regression functions the convex hull of d-fold products of functions, for example, the class of all$d$-dimensional distribution functions. For design on a grid, the exponent W does not depend on d. We connect the results to entropy estimates and show that they can be sharp. The results can also be applied to density estimation problems. When the class of densities is a mixture of a d-fold product of densities in a parametric class, the entropy of the class depends on d only in the logarithmic terms.

Thursday 29.08.2024 17:40 - 18:30 Room: Auditorioum 4 Chair: Sonja Greven Keynote talk II

Quantifying uncertainty with Bayesian deep learning

Speaker: **Nadja Klein, Karlsruhe Institute of Technology, Germany**

Bayesian deep learning fuses deep neural networks with Bayesian techniques to enable uncertainty quantification and enhance the robustness in complex tasks such as image recognition or natural language processing. However, fully Bayesian estimation for neural networks is computationally intensive, requiring the use of approximating inference for virtually all practically relevant problems. Even for partially Bayesian neural networks, there is often a lack of clarity on how to adapt Bayesian principles to deep learning tasks, leaving practitioners overwhelmed by the theoretical aspects, such as choosing appropriate priors. So, how are scalable, reliable, and robust approximate Bayesian methods designed for deep learning? The question is addressed from a methodological perspective with a focus on Bayesian neural networks, Bayesian optimization, and probabilistic programming. Methods are developed that deliver high-accuracy predictions and offer calibrated probabilistic confidence measures in those predictions. This is showcased through examples from different domains and selected open challenges and directions for future research are concluded.

Friday 30.08.2024 12:10 - 13:00 Room: Auditorioum 4 Chair: Maria Brigida Ferraro Keynote talk III

An extended latent factor framework for ill-posed generalised linear regression

Speaker: **Tatyana Krivobokova, University of Vienna, Austria** Gianluca Finocchio

The classical latent factor model for (generalised) ill-posed linear regression is extended by assuming that, up to an unknown orthogonal transformation, the features consist of subsets that are relevant and irrelevant to the response. Furthermore, a joint low-dimensionality is imposed only on the relevant features and the response variable. This framework not only allows for a comprehensive study of the partial-least-squares (PLS) algorithm under random design, but also sheds light on the performance of other regularisation methods that exploit sparsity or unsupervised projection. Moreover, we propose a novel iteratively-reweighted-partial-least-squares (IRPLS) algorithm for ill-posed generalised linear models and obtain its convergence rates working in the suggested framework.

Tuesday 27.08.2024 11:00 - 12:30 Parallel Session B – COMPSTAT2024

CI006 Room 45 RECENT ADVANCES IN OPTIMAL DESIGN OF EXPERIMENTS Chair: Stefanie Biedermann

C0292: **Computing constrained optimal designs with applications to dose-finding**
Presenter: **Lenka Filova**, Comenius University in Bratislava, Slovakia
Co-authors: Radoslav Harman, Pal Somogyi

The utility of mathematical programming methods in computing optimal designs for clinical trials is demonstrated. These methods offer a flexible approach that is particularly beneficial for dose finding, as they readily incorporate constraints that are naturally inherent to the problem. An important application of this approach is in computing optimal designs for multivariate observations. This is crucial in various scenarios, such as in cases where efficacy and toxicity are jointly investigated or when monitoring multiple indicators. As an illustration, we focus on computing locally optimal and adaptive designs in a multinomial logistic model, particularly relevant to clinical trials. We achieve this using algorithms from the R library OptimalDesign, which enables users to compute both approximate and exact designs based on several frequently used optimality criteria.

C0198: **Optimizing the allocation of trials to sub-regions in multi-environment crop variety testing for correlated genotypes**
Presenter: **Maryna Prus**, Hohenheim University, Germany

New crop varieties are extensively tested in multi-environment trials in order to obtain a solid basis for recommendations to farmers. When the target population of environments is large, a division into sub-regions is often advantageous. If the same set of genotypes is tested in each of the sub-regions, a linear mixed model (LMM) may be fitted with random genotype-within-sub-region effects. The first analytical results for optimizing the allocation of trials (designs) to sub-regions have been obtained recently. The genotype effects were assumed to be uncorrelated. However, this assumption is not always suitable for practical situations. In praxis, genetic markers are often used in plant breeding to determine the genetic relationships of genotypes, which helps to model their correlation. A more general LMM with correlated genotype effects is considered. An analytical solution for the allocation of trials is proposed in the form of an optimality condition. For particular covariance structures of genotype effects, optimal designs are given explicitly.

C0375: **Optimal designs for state estimation in networks**
Presenter: **Kirsten Schorning**, Technical University Dortmund, Germany

A design problem is addressed, which is motivated by the study of electrical power distribution grids at medium and low-voltage levels. In a specific distribution grid, the question arises where measurements of the electrical power should be taken and how precise these measurements should be in order to get a precise estimation of the state of the grid. Due to high costs, it is not possible to use sensors to measure the electrical power at each position of the grid, and at some positions, so-called pseudo measurements have to be used instead. In order to solve this design problem, two models are considered for estimating the expected states of nodes in networks where the observations at nodes are given by random states and measurement errors. In the first model, independent successive observations at the nodes are assumed, and the design question is how often the nodes should be observed to obtain a precise estimation of the expected states. In the second model, which is more realistic in the context of electrical power distribution grids, all nodes are observed simultaneously, and the design question is to determine the nodes which need larger precision of the measurements than other nodes. It is shown that both models lead to the same design problem, and A-optimal designs are derived explicitly for simple networks that have star or wheel configurations.

CO124 Room 052 NONPARAMETRIC INFERENCE AND INVERSE PROBLEMS Chair: Fabian Dunker

C0400: **A mollification approach to stabilize econometrics models: Application to deconvolution and random coefficients models**
Presenter: **Anne Vanhems**, TBS Education, France

Mollification is used to regularize ill-posed econometrics problems such as deconvolution, instrumental variables regression or random coefficients regression. This regularization method offers a unifying and generalizing framework and is particularly appealing with Fourier transform or Radon transform. Convergence properties of nonparametric regularized estimators are studied, and finite sample properties are analyzed using simulations.

C0345: **Nonparametric estimation under Gaussian measurement error with conditionally heteroscedastic variances**
Presenter: **Alexander Meister**, University of Rostock, Germany

The problem of estimating a density based on replicated observations that are contaminated by centered Gaussian noise is considered, where the conditional noise variance may depend on the unobserved random variable with the target density. Standard Fourier techniques for deconvolution seem inappropriate in the underlying setting. We introduce nonparametric estimators of both the target density and the variance function based on higher-order Kronecker moments and show that these procedures attain optimal minimax convergence rates. An extension to the corresponding errors-in-variable regression model is provided.

C0268: **From small scales to large scales: Distance-to-measure density based geometric analysis of complex data**
Presenter: **Katharina Proksch**, University of Twente, Netherlands
Co-authors: Christoph Weitkamp, Thomas Staudt, Christophe Zimmer, Benoit Lelandais

The analysis and classification of complex point clouds are considered. We focus on the task of identifying differences between noisy point clouds based on small-scale characteristics while disregarding large-scale information. We propose an approach based on a transformation of the data via the so-called Distance-to-Measure (DTM) function, a transformation which is based on the average of nearest neighbour distances. For each data set, we estimate the probability density of average local distances of all data points and use the estimated densities for classification. While the applicability is immediate and the practical performance of the proposed methodology is very good, the theoretical study of the density estimators is quite challenging, as they are based on i.i.d. observations that have been obtained via a complicated transformation. In fact, the transformed data are stochastically dependent in a non-local way that is not captured by commonly considered dependence measures. Nonetheless, we show that the asymptotic behaviour of the density estimator is driven by a kernel density estimator of certain i.i.d. random variables by using theoretical properties of U-statistics, which allows us to handle dependencies. We show via a numerical study and in an application to simulated single molecule localization microscopy data of chromatin fibers that unsupervised classification tasks based on estimated DTM-densities achieve excellent separation results.

C0297: **Doubly robust Bayesian Difference-in-Differences estimators**
Presenter: **Christoph Breunig**, University of Bonn, Germany

A double robust Bayesian inference procedure is proposed for estimating the average treatment effect on the treated (ATT) within the difference-in-differences research design. Our robustification of the Bayesian procedure involves two important modifications: first, adjusting the prior distributions of the conditional mean function, and second, correcting the posterior distribution of the resulting ATT. We prove the asymptotic equivalence between our Bayesian estimator and efficient frequentist estimators by establishing a new semiparametric Bernstein-von Mises theorem under double robustness. That is, the lack of smoothness in conditional mean functions can be compensated for by the regularity of the propensity score and vice versa. Consequently, the Bayesian credible sets form confidence intervals with asymptotically exact coverage probability. In simulations, our robust Bayesian procedure leads to a significant reduction in bias for point estimation and accurate coverage of confidence intervals, especially when the dimensionality of covariates is large relative to the sample size and the underlying functions become complex.

CO082 Room 44 TEXT MINING **Chair: Philipp Adaemmer**

C0162: **Going viral: Inflation narratives and the macroeconomy**
Presenter: **Max Weinig**, Universitat Hamburg, Germany
Co-authors: Ulrich Fritsche

For the first time in decades, many countries are facing high inflation rates. While some inflation might be desirable, rates way above the target impose costs on society. According to macroeconomic theory, inflation expectations are considered a potential driver of inflation dynamics. Concurrently, there is a growing interest in the analysis of narratives in macroeconomic research. This has created a link to modern social and psychological analysis of expectations. In the context of the current period of inflation, the relevance of narratives lies in their potential to influence the expectations of households and economic decision-makers. Contributing to this research, an approach to extract identified economic narratives from media reports is proposed. Therefore, results from a survey study on inflation narratives are combined with a keyword-assisted topic model (keyATM) and a semisupervised semantic scaling technique (LSS) to measure inflation narratives as tone-adjusted time series. To further investigate the inflation expectations' determinants, multivariate Granger causality tests and local projections are applied. The empirical analysis indicates that inflation narratives are a potential driver of inflation expectations, particularly short-run expectations.

C0189: **PETapter: A masked-language-modeling classification head for modular fine-tuning of (large) language models**
Presenter: **Jonas Rieger**, TU Dortmund University, Germany

A significant portion of applications for large language models involve classification tasks for documents, sequences, sentences, or even single entities. For such tasks, pretrained encoder-only models like RoBERTa and DeBERTa serve as powerful tools. These models typically undergo supervised fine-tuning adding an additional linear layer to the transformer architecture, known as a classification head, using datasets of varying sizes. This fine-tuning may incorporate a technique called parameter-efficient fine-tuning (PEFT), which freezes large parts of the base model to reduce computational demand. Additionally, few-shot learning methods, such as pattern-exploiting training (PET), enable faster adaptation to (few) training examples. PETapter represents a fusion of these two promising research directions, leveraging the strengths of both to achieve effective training and performant predictions with just a few training samples. It employs PEFT methods for fine-tuning word embeddings and a PET-like masked-language-modeling objective for the final classification of text elements. Utilizing a benchmark study across various datasets, we demonstrate that PETapter is computationally more efficient than full fine-tuning via PET while maintaining comparable performance with just 100 training examples. Furthermore, it surpasses the performance of classical PEFT methods when used in conjunction with traditional classification heads.

C0190: **Risky news and credit market sentiment**
Presenter: **Paul Labonne**, BI Business School, Norway
Co-authors: Leif Anders Thorsrud

The nonlinear nexus between financial conditions indicators and the conditional distribution of GDP growth has recently been challenged. We show how one can use textual economic news combined with a shallow Neural Network to construct an alternative financial indicator based on word embeddings. By design, the index associates growth-at-risk to news about credit, leverage and funding, and we document that the proposed indicator is particularly informative about the lower left tail of the GDP distribution and delivers significantly better out-of-sample density forecasts than commonly used alternatives. Speaking of theories on endogenous information choice and credit-market sentiment, we further document that the news-based index likely carries information about beliefs rather than fundamentals.

C0334: **Exploring the predictive capacity of ESG sentiment on official ratings: A few-shot learning perspective**
Presenter: **Elena Toenjes**, Justus-Liebig-University Giessen, Germany
Co-authors: Christoph Funk, Christian Haas

Environmental, social, and governance (ESG) criteria are increasingly central to corporate reporting. Natural language processing (NLP) techniques, specifically a RoBERTa-based few-shot model, are applied to conduct aspect-based sentiment analysis (ABSA). The analysis targets ESG-related entities and their sentiments within EURO STOXX 50 company reports, mapping them to an ESG sub-category to assess their impact on ESG ratings. Ratings data are sourced from established providers, including Refinitiv, Standard & Poor's, and potentially Bloomberg. Furthermore, to explore potential reciprocal influences on these variables, a panel vector auto-regressive (PVAR) model is employed, which facilitates the modeling of bidirectional interactions. The combination of advanced NLP methods and comprehensive data integration aims to provide detailed insights into the dynamics between company disclosures and rating providers' ESG scores.

CC133 Room 050 FINANCIAL TIME SERIES **Chair: Alessandra Amendola**

C0441: **Portfolio optimization using hybrid robust time series clustering and robust mean-variance portfolio selection**
Presenter: **Dedi Rosadi**, Universitas Gadjah Mada, Indonesia, Indonesia
Co-authors: Peter Filzmoser, La Gubu, Lestari Vemmie Nastiti

A novel portfolio optimization approach is presented by applying several hybrid approaches between robust clustering and robust portfolio selection approaches. When there are many stocks that can be selected during the portfolio optimization process, this approach can be used to quickly obtain the optimum portfolio, where, at the same time, the method is also robust to the presence of outliers in the data. The daily closing price of stocks listed on the Indonesia Stock Exchange, which are included in the LQ-45 indexed from August 2017 to July 2018, was used as a case study. The empirical study showed that portfolios constructed using PAM time series clustering with autocorrelation dissimilarity and a robust FMCD-MV portfolio model outperformed portfolios created using other considered approaches.

C0479: **Which early warning signals predict high-frequency extreme price movements?**
Presenter: **Philippe Hubner**, HEC Liege, University of Liege, Belgium
Co-authors: Julien Hambuckers

The dynamic distribution of block maxima time-series is modeled, with an application to high-frequency stock returns. To do so, an autoregressive structure is considered in the parameters of the generalized extreme value (GEV) distribution, which are also conditioned by past information. The recently developed penalization technique is used to select relevant covariates. In a simulation study, the finite sample properties of the estimation are inspected, and the ability to select active covariates and related computational issues is addressed. As an empirical illustration, these techniques are applied for distributional forecasting using 5-minute returns on a sample of NASDAQ stocks, with liquidity measures among the candidate covariates.

C0391: **Wald-type test for conditional moving average unit root**
Presenter: **Ryota Yabe**, Shinshu University, Japan

A Wald-type test is introduced for the unit root moving average model of order 1 (MA(1)). The MA model with a unit root is a widely studied time series model, and its unit root test is essential in various applications, such as testing the stationarity of AR processes and examining cointegration. The LM type test is very commonly used but has the drawback that its limiting power is not very high for distant alternative hypotheses. The KPSS test, which is based on this test, also exhibits the same limitation. A new Wald-type test derived from the auxiliary equation is introduced.

This test was initially proposed in a prior study for a long memory process. This test is particularly useful for complex models where the limiting distribution of the estimator of the parameter of interest is unclear. The implementation of this test can be achieved by utilizing an auxiliary equation in conjunction with an OLS estimator. The results indicate that in both finite and infinite samples, the performance of our proposed test surpasses that of the LM test. Furthermore, the limiting power function of our test is consistently near the limiting power envelope function.

C0188: **Combining caterpillar-SSA methods and mixed frequency data regression for inflation forecasting**
Presenter: **Elena Zarova**, Tashkent State University of Economics, Uzbekistan

Obtaining reliable inflation forecasts is important for any type of economy and level of economic development. Officially published monthly consumer price indices lag behind the actual market situation for a period of 1-2 months, which is critical for decision-making in the financial and real sectors of the economy, as well as for the competent economic behavior of households. Reducing this gap as much as possible is very important in conditions of economic instability. A possible approach to solving this problem, which has scientific novelty, is based on a combination of the Caterpillar - SSA (Singular Spectrum Analysis) method and the MIDASR (Mixed Frequency Data Sampling Regression Models) method. The results of multivariate forecasting of weekly consumer price indices for individual goods using the SSA method are the input for constructing a regression model of the monthly CPI for food products using MIDASR methods. Based on statistical information criteria, it is concluded that this method provides a more reliable leading estimate of the CPI relative to other multifactor modeling methods. Practical examples of solving this problem using data from a number of countries and R packages are given. The proposed method has practical importance for many forecasting problems.

CC047 Room 051 SPATIAL STATISTICS Chair: Mattias Villani

C0427: **Iterative methods for Vecchia-Laplace approximations for latent Gaussian process models**
Presenter: **Pascal Kuendig**, Lucerne University of Applied Sciences and Arts, Switzerland
Co-authors: Fabio Sigrist

Latent Gaussian process (GP) models are flexible probabilistic non-parametric function models. Vecchia approximations are accurate approximations for GPs to overcome computational bottlenecks for large data, and the Laplace approximation is a fast method with asymptotic convergence guarantees to approximate marginal likelihoods and posterior predictive distributions for non-Gaussian likelihoods. Unfortunately, the computational complexity of combined Vecchia-Laplace approximations grows faster than linearly in the sample size when used in combination with direct solver methods such as the Cholesky decomposition. Computations with Vecchia-Laplace approximations can thus become prohibitively slow precisely when the approximations are usually the most accurate, i.e., on large data sets. Iterative methods are presented to overcome this drawback. Among other things, several preconditioners are introduced and analyzed, new convergence results are derived, and novel methods are proposed for accurately approximating predictive variances. The proposed methods are analyzed theoretically and in experiments with simulated and real-world data. In particular, a speed-up of an order of magnitude compared to Cholesky-based calculations and a threefold increase in prediction accuracy in terms of the continuous ranked probability score compared to a state-of-the-art method on a large satellite data set are obtained.

C0485: **Flexible inference for spatiotemporal Hawkes processes with general parametric kernels**
Presenter: **Emilia Siviero**, Telecom Paris, France
Co-authors: Guillaume Staerman, Stephan Clemencon, Thomas Moreau

With advancements in data collection technologies, fields such as sociology, epidemiology, and seismology are increasingly encountering spatiotemporal datasets with self-exciting properties characterized by triggering and clustering behaviors that can be effectively modeled using a Hawkes space-time process. A fast and flexible parametric inference method is developed to estimate the parameters of the kernel functions in the intensity function of a space-time Hawkes process based on such data. The statistical approach integrates three main components: 1) the use of kernels with finite support, 2) appropriate discretization of the space-time domain, and 3) the use of (approximate) precomputations. The proposed inference technique employs an ℓ_2 gradient-based solver, which is both fast and statistically accurate. In addition to describing the algorithmic aspects, numerical experiments have been carried out on synthetic and real spatiotemporal data, providing solid empirical evidence of the relevance of the proposed methodology.

C0394: **Estimation of spatiotemporal extremes via generative neural networks**
Presenter: **Christopher Buelte**, Ludwig-Maximilians-Universitat Munchen, Germany
Co-authors: Lisa Leimenstoll, Melanie Schienle

Recent methods in modeling spatial extreme events have focused on utilizing parametric max-stable processes and their underlying dependence structure. A unified approach is provided for analyzing spatial extremes with little available data by estimating the distribution of model parameters or the spatial dependence directly. By employing recent developments in generative neural networks, a full sample-based distribution is predicted, allowing for direct assessment of uncertainty regarding model parameters or other parameter-dependent functionals. The method is validated by fitting several simulated max-stable processes, showing a high accuracy of the approach regarding parameter estimation, as well as uncertainty quantification. Additional robustness checks highlight the generalization and extrapolation capabilities of the model, while an application to precipitation extremes across Western Germany demonstrates the usability of the approach in real-world scenarios.

C0417: **Deep parametric predictive Gaussian processes for uncertainty estimation**
Presenter: **Oluwole Oyebamiji**, University of Birmingham, United Kingdom

Deep Gaussian processes (DGPs) are a powerful extension of Gaussian processes that allow for multi-layer generalization of GPs, enabling more flexible and expressive modelling of complex data. However, as the depth of the model increases, so does the computational cost, making it challenging to scale deep Gaussian processes to large-dimensional data. This often leads to underestimation of the posterior variance. Moreover, interpreting and understanding the learned representations in DGPs can be more difficult than in shallower models. The model developed combines a hybrid spatial factor model that reduces the difficulty of dealing directly with high-dimensional outcomes and a Bayesian method that integrates input variability into GP regression. The proposed model used inducing point methods with stochastic variational inference, which provides substantially improved predictive uncertainties and efficient approximation. The benefits of the model are evaluated on several benchmark regression datasets and high-dimensional data from the IMPRESSIONS integrated assessment platform, version 2. The performance of the input features is analyzed using the proposed models and SHapley Additive exPlanations (SHAP) values for multi-task problems to help interpret the results. The results show that the proposed integration of these techniques is efficient and accurate for the uncertainty quantification of complex models.

CC051 Room 43 FUNCTIONAL DATA ANALYSIS Chair: Sonja Greven

C0378: **Density-on-scalar regression for bivariate distributions in Bayes spaces**
Presenter: **Ivana Pavlu**, Palacky University Olomouc, Czech Republic
Co-authors: Almond Stoecker, Adela Czolkova, Karel Hron, Sonja Greven

Additive regression models allow the explanation of the dependence of a response variable on a set of covariates through a parameterized yet flexible regression model. Recent advances allow functional and, more recently also, distributional responses. This contribution presents additive density-on-scalar regression for bivariate distributional responses, either continuous (probability density functions) or discrete (probability distribution/compositional tables). The regression model, embedded in the Bayes space theory, enables a sound analysis of the relationship between the response and covariates, respecting the relative properties of the distributional response. One of the challenges here is to find a good means of interpretation of estimated effects. To achieve this, the model allows the decomposition of bivariate density-on-scalar effects by 1. Orthogonal decomposition of additive effects into linear, non-linear and covariate-interaction parts, and 2. Orthogonal decomposition of the bivariate distribution into an interactive part and two (geometric) margins. Due to this second decomposition, model selection -performed through gradient boosting - allows assessment of the importance of different covariates and independence between the two response variables in their mean distribution and/or covariate effects. The proposed model framework is used for the analysis of salary and workload distributions of cohabiting couples based on the German Socio-Economic Panel study.

C0450: **Dealing with count zeros in preprocessing of probability density functions**

Presenter: **Stanislav Skorna**, Palacky University, Czech Republic

Co-authors: Jitka Machalova, Karel Hron

Probability density functions occur naturally as data in applications in many different fields, e.g. income distribution in econometrics, mortality densities in epidemiology or soil contamination in geochemistry. However, since densities are mostly unobserved in their continuous form and available only through discrete samples, their reliable approximation is crucial for employing modern methods of functional data analysis. Zero counts or missing values are often presented in the samples, which causes the main issue in the preprocessing, e.g. for using centered log-ratio transformation. The contribution opens a discussion about the issue and presents possible solutions using adaptations of imputation methods for (multivariate) compositional data and sparse functional data analysis illustrated on simulated and real data.

C0480: **Functional PARAFAC with probabilistic modeling**

Presenter: **Lucas Sort**, CentraleSupelec, France

Co-authors: Laurent Le Brusquet, Arthur Tenenhaus

In longitudinal studies, data can increasingly be organized as tensors. Within this framework, the time-continuous property often implies a smooth functional structure on one of the tensor dimensions. To help researchers investigate such data, a new tensor decomposition approach is introduced based on the CANDECOMP/PARAFAC decomposition. The approach allows for the representation of a high-dimensional functional tensor as a low-dimensional set of functions and feature matrices. Furthermore, to capture the underlying randomness of the sampling setting more efficiently, a probabilistic model is introduced in the decomposition. A block-relaxation algorithm using only covariance and cross-covariance operators is derived to obtain estimates of model parameters. Thanks to the covariance formulation of the solving procedure and the probabilistic modeling, the method can be used in sparse and irregular sampling schemes, making it applicable in numerous settings. Intensive simulations are introduced to show the notable advantage of the method in reconstructing tensors and retrieving insightful latent information.

C0484: **Function-on-scalar regression via first-order gradient-based optimization**

Presenter: **Quentin Edward Seifert**, Georg-August-Universitaet Goettingen, Germany

Co-authors: Elisabeth Bergherr, Tobias Hepp

Functional regression models allow for the inclusion of functional covariates and responses. Due to the nature of the data, the models can quickly become computationally expensive, even with a comparably low number of observations. To address this increased complexity, gradient descent-based functional regression is introduced. The idea is to fit functional regression models using gradient descent-based optimization algorithms and estimate the model parameters as one would estimate the parameters of neural networks. The proposed model provides an easily scalable and customizable alternative to established approaches. Preliminary simulation results show that the approach performs reliably. The approach is applied to supermarket parking data recorded during the first months of the Covid-19 pandemic in Germany to analyze the effect of the contact restrictions introduced during this period on consumer behavior.

Tuesday 27.08.2024 14:00 - 15:30 Parallel Session C – COMPSTAT2024

CI005 Room 44 ADVANCED STATISTICS IN RISK MODELLING AND HIGH-DIMENSIONAL ESTIMATION Chair: Alessandra Amendola

C0347: **Measuring climate-related and environmental risks for equities**
Presenter: **Emese Lazar**, University of Reading, United Kingdom
Co-authors: Shixuan Wang, Jingqi Pan

Financial regulators and investors are increasingly concerned about the effects of climate change on investments and seek to capture the climate-related and environmental risks of investments. While energy companies have attracted most of the attention due to the contribution of the energy sector to environmental degradation, climate-related and environmental risks actually affect companies in every sector. Novel measures termed climate value-at-risk (VaR) and climate expected shortfall (ES) are proposed to capture the risk attributed to transition risk factors proxied by environmental scores. The average ratio of climate VaR and ES to total risk in various equity sectors are compared, which enables the identification of the sectors in which climate and environmental risk factors contribute most to the total risk. The analysis considers different risk measurements and various significance levels. Findings show the heterogeneity in sensitivity to climate and environmental risk factors in various sectors. The healthcare sector is the least cost-effective in reducing climate-related and environmental risks, and the energy sector benefits most from improving the firm's environmental scores.

C0315: **Recent advances in high dimensional estimation of diffusion models**
Presenter: **Mark Podolskij**, University of Luxembourg, Luxembourg

Recent statistical results for high-dimensional problems in the context of diffusion processes will be reviewed. Such models are frequently used in financial applications, among other fields, and we will investigate high-dimensional estimation of the drift and volatility matrix.

C0167: **(Quantile) Spillover indexes: Simulation-based evidence, confidence intervals and a decomposition**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Giovanni Bonaccolto, Jawad Shahzad

Quantile-spillover indexes have recently become popular for analysing tail interdependence. In an extensive simulation study, we show that the estimation of spillover indexes is affected by a positive distortion when the parameters of the underlying fitted models are not evaluated in terms of their statistical significance. The distortion is reduced for increasing sample sizes, thanks to the consistency of estimators, or by filtering out non-significant parameters. Even if in small samples, it does not fully disappear due to type I error. We make another step by introducing a simulation-based approach to recovering confidence intervals from quantile spillover indexes. In addition, we put forward an algebraic decomposition of quantile spillover separating the dynamic interdependence from the contemporaneous interdependence (due to residual correlation). Empirical evidence shows that distortions in real data are sizable, and the decomposition points out that most of the spillover is due to contemporaneous effects. All of our results extend and are confirmed for the Spillover index.

CO084 Room 43 ADVANCES IN FUNCTIONAL AND COMPLEX DATA ANALYSIS Chair: Bo Wang

C0283: **Interval estimation for continuous-time correlation**
Presenter: **Philip Reiss**, University of Haifa, Israel
Co-authors: Biplab Paul, Noemi Foa, Dror Arbiv

Continuous-time correlation is a recently proposed way to measure association between time series that are noisy and possibly irregularly observed. The crux of the method is basis-function smoothing of the time series, which can effectively mitigate the well-known attenuation problem for correlation estimation with noisy data. This technique is reminiscent of functional data analysis, but treats the observations, rather than the variables, as forming a continuum. We focus on interval estimation for continuous-time correlation. We present two approaches, one by bootstrapping and the other by posterior simulation. The latter is faster, but only the former allows the two variables to be observed at different time points. Moving beyond inferring the correlation between two curves observed with noise, we also consider inference for the correlation parameter of the underlying bivariate stochastic process. The methods are validated by simulation, and are illustrated with incompletely observed international development data and with electroencephalography data.

C0287: **Bayesian hierarchical latent variable-based modelling for large and complex genomic datasets**
Presenter: **Mayetri Gupta**, University of Glasgow, United Kingdom

Advances in genomic sequencing technologies in the past few decades have opened up the possibility of making previously inconceivable biological discoveries at extremely high resolution- but have led to numerous challenges in accurately analysing the generated data. These data are typically of huge dimension- leading to computational obstacles; are subject to various artefacts; and their distributions exhibit complex features, such as long-ranging correlations, non-ellipsoidal shapes, skewness and multimodality, causing difficulties in inference through standard statistical models. We will discuss some recent examples of Bayesian hierarchical modelling and inference for complex genomic data, along with robust, efficient and powerful computational methods enabling inference and biological discovery. One example involves clustering non-ellipsoidal data- finding subgroups with common features is often a necessary first step in the statistical analysis of large and complex genomic datasets. Another relates to detecting differential epigenetic profiles from high-throughput sequencing data. The performance of these methods is illustrated in simulation studies and applications to real-life examples from genotyping, and DNA methylation studies. This is based on joint work with Edoardo Redivo, Hien Nguyen, Huizi Zhang, Ben Swallow, Tushar Ghosh, Vincent Macaulay and Peter Adams.

C0385: **Cross-national comparisons of Covid-19 lockdown effectiveness: The spatial functional data analysis approach**
Presenter: **Pipat Wongsa-art**, Cardiff University, United Kingdom

Although studying the cross-national effectiveness of lockdown strategies in reducing the transmission of Covid-19 is necessary and extremely important, it is far from being straightforward. The endogeneity of policy choices and reverse causality are obvious examples of obstructions that may impede progress. The problem is transformed into analyzing spatially dependent discrete longitudinal data of Covid19 cases and deaths, which are often used by governments as the basis for making policy decisions. In the context of the analysis, the spatial dependence is extended beyond the concept of physical contiguity of neighborhoods, which is often the focus in spatial econometrics, to Covid19-response contiguity. Furthermore, a novel functional data analysis approach is suggested that can help disentangle a component of the data series of Covid19 cases/deaths, which is due to the Covid19-response contiguity, from another component that is country-specific. The usefulness of the latter resides in its ability to capture information about the effectiveness of government policies. The method is used to perform cross-national comparisons of Covid-19 lockdown effectiveness in 36 OECD countries and provides a number of insights that are not yet available in the literature.

C0243: **Clusterwise nonlinear regression with Gaussian processes methods**
Presenter: **Bo Wang**, University of Leicester, United Kingdom

Clusterwise regression, also referred to as regression clustering, is a technique that combines regression analysis and cluster analysis to discover relationships within data where more than one relationship exists between response variables and explanatory variables. It aims to estimate the different relationships and partition the data points simultaneously. This problem was first introduced by Spath in 1979, and an abundance of further developments have been studied since then. However, almost all clusterwise regression models and algorithms in the literature are based on linear regression, and the nonlinear regressions are limited to the cases where a family of nonlinear functions are assumed and only the unknown parameters are to be determined, such as polynomial functions, Fourier basis functions. We consider clusterwise nonlinear regression problems based on Gaussian process regression without assuming the form of candidate nonlinear functions. A K-means-like clustering algorithm is proposed, and numerical examples demonstrate its effectiveness. The method is also extended to functional nonlinear regression with scalar response and functional and scalar predictors.

CO119 Room 45 SPATIO-TEMPORAL AND NETWORK ANALYSIS Chair: Carsten Jentsch

C0202: Matrix autoregressive model with vector time series covariates for spatio-temporal data

Presenter: **Yang Chen**, University of Michigan, United States

A new model is presented for forecasting time series data distributed on a spatial grid, using historical spatiotemporal data and auxiliary vector time series data. The historical matrix time series are mapped to the predicted matrix via row and column-specific autoregressive coefficient matrices. The vector predictors are mapped to the predicted matrix by taking a mode-product with a 3D tensor coefficient. Given the high dimensionality and underlying spatial structure, we represent the tensor coefficient by a functional-coefficient from a Reproducing Kernel Hilbert Space. We jointly estimate the autoregressive and functional coefficients under a penalized maximum likelihood estimation coupled with an alternating minimization algorithm. Large sample asymptotics of the estimators are established. The performance of the model is validated with extensive simulation studies. A real data application to forecast the total electron content maps, an important parameter for monitoring space weather impacts, will be presented.

C0227: A multivariate spatial and spatiotemporal ARCH Model

Presenter: **Philipp Otto**, University of Glasgow, United Kingdom

Multivariate spatiotemporal autoregressive conditional heteroscedasticity (ARCH) models based on a vec-representation are presented. The model includes instantaneous spatial autoregressive spill-over effects, as they are usually present in geo-referenced data. Furthermore, spatial and temporal cross-variable effects in the conditional variance are explicitly modelled. We transform the model to a multivariate spatiotemporal autoregressive model using a log-squared transformation and derive a consistent quasi-maximum-likelihood estimator (QMLE). For finite samples and different error distributions, the performance of the QMLE is analysed in a series of Monte-Carlo simulations. In addition, we illustrate the practical usage of the new model with a real-world example. We analyse the monthly real-estate price returns for three different property types in Berlin from 2002 to 2014. We find weak (instantaneous) spatial interactions, while the temporal autoregressive structure in the market risks is of higher importance. Interactions between the different property types only occur in the temporally lagged variables. Thus, we see mainly temporal volatility clusters and weak spatial volatility spillovers.

C0303: Autoregressive dynamic network modelling with serial and cross-sectional dependence

Presenter: **Jonathan Flossdorf**, TU Dortmund University, Germany

Co-authors: Daniel Dzikowski, Carsten Jentsch

A flexible dynamic network modelling approach is proposed based on a class of generalised binary vector autoregressive (gbVAR) models. Originally designed for multivariate binary data with serial and cross-sectional dependence, it is also well suited for modelling dynamic networks characterized by a time series of binary adjacency matrices. This is due to the fact that gbVAR models are parsimoniously parametrized, well interpretable, and also allow for the modelling of negative dependence. As they are autoregressive in nature and satisfy the classical Yule-Walker equations, the recently developed toolbox for penalised estimation of (continuous) vector autoregressive (VAR) models enables parameter estimation of reasonably large dynamic networks observed at moderately many time points under sparsity constraints. In this context, we consider a lasso-penalised estimation procedure that particularly allows the incorporation of additional information in the form of linear restrictions for further dimension reduction. We evaluate the proposed approach and illustrate our theoretical findings with extensive simulations. The applicability is further demonstrated by some real-world examples.

C0348: Testing for global covariate effects in dynamic interaction event networks

Presenter: **Alexander Kreiss**, Leipzig University, Germany

Co-authors: Enno Mammen, Wolfgang Polonik

In statistical network analysis, it is common to observe so-called interaction data. Such data is characterized by actors forming the vertices and interacting along the edges of the network, where edges are randomly formed and dissolved over the observation horizon. In addition, covariates are observed, and the goal is to model the impact of the covariates on the interactions. Two types of covariates are distinguished: global, system-wide covariates (i.e., covariates taking the same value for all individuals, such as seasonality) and local, dyadic covariates modeling interactions between two individuals in the network. Existing continuous-time network models are extended to allow for comparing a completely parametric model and a model that is parametric only in the local covariates but has a global nonparametric time component. This allows, for instance, testing whether global time dynamics can be explained by simple global covariates like weather, seasonality, etc. The procedure is applied to a bike-sharing network by using weather and weekdays as global covariates and distances between the bike stations as local covariates.

CC138 Room 050 CHANGE POINT ANALYSIS Chair: Roland Fried

C0222: A non-parametric method for high dimensional change point analysis

Presenter: **Lupeng Zhang**, Durham University, United Kingdom

Co-authors: Reza Drikvandi

Change point analysis aims to detect significant changes in the distribution of a data sequence. It holds critical importance across modern statistical applications such as economics, finance, quality control, genetics and medical research. While change point detection for low dimensional data is extensively studied in the literature, change point detection is very challenging in high dimensional data where the number of variables is much larger than the number of observations. High-dimensional change point analysis has become a vital focus of recent research. We discuss the main challenges and difficulties with high dimensional change points and introduce a nonparametric approach to tackle some of those challenges. The proposed method is based on some dissimilarity distances and CUSUM statistics to detect significant change points in high dimensional data. Our method can detect changes in both mean and variance of high dimensional observations, as well as other distributional changes. We present simulation results and a real data application on the S&P 500 data.

C0369: Methods for structural change detection in the trend function of random fields

Presenter: **Sheila Goerz**, TU Dortmund University, Germany

Co-authors: Roland Fried

Sudden changes in the structure of the data can occur not only in temporal data/time series but also in spatial and spatiotemporal data. In all cases, it is important that structural breaks are recognized reliably. Change point detection in spatial data is addressed. The already available methods for this type of problem are either not designed for continuous data, impose very strict assumptions or are computationally expensive. A method for detecting an arbitrary number of changes of any type is proposed in the mean of a time series that partitions the data into blocks and calculates the Ginis mean difference of the block means. This idea is extended to the detection of changes in spatial data such as satellite imagery or disease spread data. Gini's mean difference is not limited to other choices of change point statistics that are applied to the block means. The asymptotic behavior of suitable test statistics is investigated under the hypothesis of a constant mean when applied to independent spatial data. Simulation studies indicate that the tests work well not only for independent data but also for moving average-type random fields. In ongoing work, extensions to further kinds of random fields and more general spatial dependence structures are elaborated.

C0406: **A fast Bayesian online changepoint detection algorithm**
Presenter: **Ziyang Yang**, STOR-i Doctoral Training Centre, Lancaster University, United Kingdom
Co-authors: Paul Fearnhead, Idris Eckley

Changepoint detection is crucial for online monitoring in many fields, i.e., finance and environment. While frequentist algorithms excel at quickly detecting changes in real-time, there is often a need for more information, such as quantifying the uncertainty of a change or its location. Bayesian changepoint algorithms address this need by providing a posterior distribution. This posterior can be calculated analytically if independence between models is assumed before and after a changepoint and conjugate priors are used. However, the computational complexity of these approaches grows linearly over time, resulting in quadratic time complexity. To overcome this challenge, a fast Bayesian online changepoint detection algorithm is proposed. The quadratic time complexity can be reduced to a constant level by merging potential locations of changes with similar posterior distributions on the post-change parameter. The resulting posterior distribution, which has fewer support points, can still be used for inference. In simulations, the algorithm has a similar speed but higher accuracy compared to a benchmark pruning approach, which only prunes the candidate with the lowest posterior probability. Moreover, the proposed method can also be applied to a range of different models, i.e., detecting changes in slope or slope with seasonality.

C0490: **High-dimensional penalized regression for linear time series with change-point detection**
Presenter: **Soudeep Deb**, Indian Institute of Management Bangalore, India

A novel approach for analyzing high-dimensional linear time series data with potential structural breaks is proposed. The method integrates penalized regression techniques with robust change-point detection to address the challenges of high-dimensionality and temporal shifts. The concepts of L_p regularization are employed to achieve sparse and interpretable models while ensuring consistency and robustness. The change point, if it exists, is assumed to be due to a covariate threshold and estimated from the regression model itself, following the idea of a prior study. Theoretical guarantees of the estimators are provided, and their efficacy is established through extensive simulation studies under various scenarios. The approach is further validated through a real-world application to financial time series data, illustrating its practical utility in identifying shifts in market regimes. The results show that incorporating change-point detection in high-dimensional settings significantly enhances model accuracy and interpretability, offering valuable insights for time series analysis in complex environments. The contribution is to the advancement of high-dimensional statistical methodologies and provides a robust framework for future research in time series modeling and change-point analysis.

CC018 Room 051 BAYESIAN STATISTICS Chair: Mattias Villani

C0184: **FBMS: An R package for Flexible Bayesian Model Selection**
Presenter: **Florian Frommlet**, Medical University Vienna, Austria
Co-authors: Aliaksandr Hubin, Geir Olve Storvik, Jon Lachman

The R package FBMS has implemented a highly flexible approach to construct nonlinear parametric regression models. Bayesian inference is performed using a genetically modified mode jumping Markov chain Monte Carlo algorithm (GMJMCMC), which is at the same time used to hierarchically generate non-linear features. The flexibility of FBMS both pertains to the model itself as well as to the generation of non-linear predictors. With respect to the choice of models we will demonstrate how to apply FBMS for the generalized linear model and then go beyond to use pretty much any parametric model for which the likelihood function can be specified. With respect to the prediction function, the space of non-linear features which can be generated includes many familiar model families as special cases, like, for example, fractional polynomials, neural networks or logic regression. We will show how to make use of FBMS to perform variable selection for some specific families of non-linear features and illustrate the good performance of FBMS compared to some competitors.

C0445: **Nuisance parameters, modified profile likelihood and Jacobian prior**
Presenter: **Guangjie Li**, Cardiff University, United Kingdom
Co-authors: Roberto Leon-Gonzalez

In a model with nuisance parameters, the maximum likelihood estimators (MLE) of the parameters of interest can be biased. One can reduce the bias due to the presence of the nuisance parameters by removing the O(1) bias of the profile likelihood score. To achieve this, the Jacobian integrated likelihood (JIL) is proposed, obtained by using a prior consisting of the Jacobian determinant of the new nuisance parameters, which are functions of the original nuisance parameters and are independent of the dependent variable. The JIL is closely related to the modified profile likelihood (MPL) of a past study. The adjusted MPL is proposed, which is easier to compute and can also remove the O(1) bias of the profile likelihood score. For panel fixed effects models, both the JIL and the adjusted MPL can remove the bias of order O(1/T) in the MLE as the cross-sectional size (N) increases. The conditions are given when the estimators from the adjusted MPL and the JIL are the same and consistent with T=o(N). Although the adjusted MPL and the JIL do not always exist, one can use their first-order conditions to obtain bias-reduced estimators. The theoretical results are demonstrated by panel binary choice models and dynamic panel linear models with exogenous and predetermined regressors.

C0456: **Clustered variable-order Bayesian Markov models with applications in cyber-security**
Presenter: **Daniyar Ghani**, Imperial College London, United Kingdom
Co-authors: Nick Heard, Francesco Sanna Passino

Many models for categorical sequences assume exchangeable or first-order dependent sequence elements. These are common assumptions, for example, in models of computer malware traces and protein sequences. Although such simplifying assumptions lead to computational tractability, the models often fail to capture long-range, complex dependence structures that may be harnessed for greater predictive power. For example, in cyber-security, it is known that different event types on a network have strong correlations and processes that run on a computer also exhibit complex dependencies, as one process usually triggers the execution of child processes. To this end, a Bayesian modelling framework is proposed to parsimoniously capture rich dependence structures in categorical sequences, with memory efficiency suitable for real-time processing of data streams. A clustered, variable-order Markov model with conjugate prior distributions is developed. The novel framework requires fewer parameters than fixed order-n Markov models by dropping redundant dependencies and clustering sequential contexts. Approximate inference is performed via model-based clustering and Markov chain Monte Carlo methods, demonstrated on synthetic and real-world data examples. Practical applications include interpretable analysis of cyber-attack patterns and real-time next-command prediction. The proposed model outperforms existing sequence models when fitted to a novel dataset of honeypot command-line sessions.

C0469: **Bayesian learning lithium-ion open circuit voltage curve via state-space model**
Presenter: **Tomas Iesmantas**, Kaunas University of Technology, Lithuania
Co-authors: Robertas Alzbutas

The state of charge (SoC) is an indicator of the remaining battery charge and must be continuously monitored by the battery management system. However, SoC cannot be directly measured while the battery is in use. Estimating the SoC, particularly the open-circuit voltage curve, which defines the relationship between SoC and open-circuit voltage, remains a challenge. Most current techniques rely on laboratory measurements and the OCV curve derived from those measurements. However, removing the battery (e.g., from an electric vehicle) for testing is impractical for most applications. A novel application of a state-space model is presented for estimating the OCV curve based solely on current and voltage data measured during typical battery use, eliminating the need for laboratory testing. To the authors' knowledge, this is the first demonstration of estimating the OCV curve using only voltage and current measurements obtained within the context of real-world battery usage. To achieve this, the battery is modeled using an equivalent circuit model, where the SoC-OCV curve is represented by a parametric nonlinear function. The unknown parameters of the model are estimated via a Bayesian inference framework implemented using the particle MCMC algorithm. Using datasets from real batteries operating under varying workloads, the approach and its accuracy in estimating is demonstrated not only the SoC-OCV curve but also other parameters within the equivalent circuit model.

CC026 Room 052 APPLIED STATISTICS AND DATA ANALYSIS Chair: Stefanie Biedermann

C0419: **Joint models for longitudinal and time-to-event data in the social sciences**

Presenter: **Sophie Potts**, University of Goettingen, Germany
Co-authors: Karin Kurz, Anja Rappl, Elisabeth Bergherr

As joint models for longitudinal and time-to-event data (JM) are a well-established estimation method in biostatistics but do not belong to the standard toolkit of social scientists, the analysis demonstrates its usage and usefulness for an application on marital satisfaction and time to marriage dissolution. The advantages of JMs for social science research questions are highlighted, and the results are compared with classical approaches such as a time-varying covariate (TVC) and the two-stage model. With the separate JMs by gender, the expected negative current value association for marital satisfaction and the risk of marriage dissolution is found. Using a classical TVC model, the effect of marital satisfaction on the risk of marriage dissolution is highly underestimated. Furthermore, applying the JM allows the decomposition of the effect of shared household work into an insignificant *direct* effect and a highly significant *indirect* effect via marital satisfaction for women. The decomposition for men results in both effects being significant, i.e. that a higher share of household work for men is associated with a higher risk of marriage dissolution through both pathways, directly and indirectly via marital satisfaction. The models control for the standard socio-economic variables, premarital cohabitation, and children, as well as for gender-role attitudes.

C0461: **Group anomaly detection for optimizing urban planning of rental bike services**

Presenter: **Lixuan An**, Ghent University, Belgium
Co-authors: Bernard De Baets, Stijn Luca

In major cities, bike-sharing programs provide a convenient and eco-friendly transportation mode. However, managing and maintaining a large fleet of rental bikes can be logistically challenging and costly. Rental bike rides in Munich over the past five years (2019-2023) from the Munchner Verkehrsgesellschaft (MVG) bike-sharing service are analyzed to optimize the spatial arrangement of rental bike stations and free return regions through urban planning initiatives. Urban planning tasks are solved through the point process model of extreme value theory, a group anomaly detection technique. To identify potential free return regions in non-free return areas, a group anomaly detection task is built based on bike ride end locations. All bike rides ending in a specific bike station region in a non-free return area form a group, where the expected distance from the end location to the nearest station should be close to zero. In this setting, anomalous groups might indicate potential free return regions for urban planning. Furthermore, another group anomaly detection task is aimed at optimizing the distribution of bike stations. In this case, a group refers to all bike rides starting from the same bike station and ending in non-free return areas. Anomalous groups provide valuable insights for improving the distribution of station locations, ensuring better accessibility and convenience for users.

C0181: **Coherent forecast and criminal justice program evaluation in hierarchical time series**

Presenter: **Thomas Fung**, Macquarie University, Australia
Co-authors: Joanna Wang

Crime time series data can often be naturally disaggregated based on various attributes of interest, such as crime type or geographical location. When modelling this type of data, the current practice in crime science is to model each series at the most disaggregated level, as it helps to identify more subtle changes. However, authorities and stakeholders often focus on the bigger picture, leading researchers to either simply sum the fitted value series or model the aggregated series independently. This practice often leads to poorer performance at the higher levels of aggregation as the most disaggregated series typically exhibit a high degree of volatility, while the most aggregated series tends to be smoother and less noisy. We will demonstrate how the hierarchical and grouped time series structure can be utilised to provide coherent estimates for all disaggregate and aggregate series while also reconciling them to enhance forecast and criminal justice program evaluation by using all the available information. We will utilise NSW and US crime data alongside the COVID lockdown as the intervention effect for illustrative purposes.

C0429: **Spatial confounding in gradient boosting**

Presenter: **Lars Knieper**, Georg-August-University of Goettingen, Germany
Co-authors: Thomas Kneib, Elisabeth Bergherr

'Inside Airbnb' offers high-dimensional data sets of Airbnb accommodations, which include variables with information about the accommodations as well as their coordinates. There is a clear trend to higher prices in city centers, even though it might not be completely the center itself causing higher prices but the proximity to sights and alike. Hence, when complementing 'Inside Airbnb' data with covariates that describe the location, these covariates are correlated with the spatial trend of prices. This collinearity between covariates and spatial effects can lead to a bias in the corresponding fixed effects estimates, known as spatial confounding. Recently, the Spatial+ approach suggests regressing the spatial effect in the covariate first before estimating the model of interest. Drastic spatial confounding is observed in gradient boosting due to its step-wise procedure. The suggested two-step approach is applied, and its ability to correct spatial confounding for gradient boosting is also confirmed. A major advantage of this correction approach is that the gradient boosting algorithm itself does not need an alteration as the correction happens beforehand. Additionally, correcting also non-confounded covariates does not decrease estimation performance.

Tuesday 27.08.2024 16:00 - 18:00 Parallel Session D – COMPSTAT2024

CO118 Room 052 STATISTICAL METHODS FOR ENERGY AND TRANSPORT DATA Chair: Roland Fried

C0229: **Statistics of the power grid frequency**
Presenter: **Dirk Witthaut**, Forschungszentrum Juelich GmbH, Germany

A reliable supply of electric power is vital for our society. The stable operation of the power system requires that generation and load are balanced, which is ensured by elaborate control systems. Temporary imbalances occur, both due to rapid stochastic fluctuations and slower effects of scheduling and electricity trading. Such imbalances manifest in the grid frequency, which is constantly monitored and used to control generation to restore the balance. We discuss the intricate statistical properties of frequency recordings from various grids around the world with a focus on the emergence of heavy tails and long correlations. We introduce a physics-inspired machine learning model that bridges time scales from seconds to hours: Stochastic differential equations describe the fast dynamics of the frequency and the control system, while artificial neural networks are used to incorporate external influences such as scheduling and trading. The model successfully reproduces important statistical properties of the frequency in the European grid and enables probabilistic forecasting.

C0343: **Probabilistic forecasting and reconciliation of wind turbine power**
Presenter: **Antonia Arsova**, TU Dortmund University, Germany
Co-authors: Sven Pappert

New models are explored for probabilistic forecasting of hierarchical time series with an application to wind turbine power production. Considering different levels of (cross-sectional) aggregation of individual time series may yield a high-dimensional hierarchy where forecasts at each aggregation level are required. A desirable property of such forecasts is that they obey the same linear restrictions prescribed by the hierarchy as the original time series: e.g., the sum of wind power forecasts for all districts in a given region has to sum up to the forecast for that whole region. One way to achieve this coherency is forecast reconciliation. Base probabilistic forecasts for the individual time series are obtained using the classical ARIMA-GARCH model and the recently introduced MAGMAR-Copula model. Existing approaches (score optimal and bottom-up) to reconcile the base forecasts are compared, and the results are evaluated by the CRPS and the energy score. Furthermore, new reconciliation approaches are explored by incorporating nonlinear instead of linear transformations of the base forecasts to form the reconciled ones. The question of optimality of reconciliation methods for probabilistic forecasts is also explored, and different distance-based criteria are considered.

C0364: **Probabilistic forecasting of energy time series with diffusion models**
Presenter: **Nicole Ludwig**, University of Tubingen, Germany

Energy systems are complex systems with multiple time series, such as available solar and wind power, interacting with external factors such as the weather and human behaviour. Probabilistic forecasting of these systems is crucial for managing the electricity network, acting upon imbalances and planning future energy system expansion. Recent advances in machine learning, such as transformers and diffusion models, show excellent capabilities in forecasting time series. However, in energy-related forecasting tasks, where seasonality and (future) covariates play an essential role, they rarely provide an additional benefit. The aim is to investigate how diffusion models can be enhanced to outperform simpler methods, especially when including complex future weather covariates and simultaneously forecasting multiple correlated time series. A particular focus is on the uncertainty estimates constructed by the different models, especially their marginal and conditional calibration and their ability to properly propagate the uncertainty from the weather input to the power output in the energy system. The base probabilistic forecasting performance regarding the calibration is compared, and statistical post-processing is assessed to see potential performance increases in the models post hoc.

C0319: **Statistical methods for power demand and consumption time series at household level in Mexico**
Presenter: **Jorge Gonzalez-Ordiano**, Universidad Iberoamericana, A.C., Mexico
Co-authors: Milagros Santos-Moreno, Karla Obermeier-Velazquez, Luis Cortes-Munoz, Jose Asse-Amiga, Luis Corral-Corona

The risks associated with climate change have made the energy transition necessary. Nevertheless, transitioning to a more sustainable energy system raises several issues. For instance, the volatility of some renewable power sources can lead to power supply and demand fluctuations, making it difficult to maintain grid stability and reliability. Smart grids offer a promising solution to address these challenges thanks to the integration of advanced information and communication technologies, such as innovative machine learning-based forecasting models and control algorithms. However, the development of the Smart Grid is not without its challenges. For instance, in Mexico, the availability of energy-related data, particularly at the household level or from non-large urban areas, is often limited, thus complicating the development of Smart Grid technologies fine-tuned to the country's reality. For this reason, the first part of the talk will present the results of a project that collected consumption and demand time series from 5 households within a small community in Mexico at various resolutions. Afterwards, the talk will focus on the statistical methods being developed using the previously collected data to preprocess and forecast the time series and identify and analyze the possible demand flexibility.

C0470: **Analyzing the route choice of cyclists using machine learning models**
Presenter: **Katrin Lubashevsky**, TUD Dresden University of Technology, Germany
Co-authors: Iryna Okhrin, Stefan Huber, Sven Lissner

Cycling is a crucial part of the transition to a more climate-friendly transportation system. It is therefore advisable to promote cycling in transport planning, where an understanding of the influences on the choice of cycling route is necessary for efficient planning. Using GPS tracks of cycling trips and additional route information, the route choice can be modeled in various ways. To date, logit models have been the predominant type of models used in this domain. The present contribution employs a variety of machine learning techniques, including neural networks, support vector machines, decision trees, and random forests, to address the route choice problem. Data from the Germany-wide "City Cycling" campaign is used for this purpose, with the city of Freiburg, in particular, being the subject. A total of 418,620 bicycle trips were recorded for the city of Freiburg. The resulting models are evaluated and compared according to their interpretability (using partial dependence plots and variable importance plots) and predictive quality (using metrics such as recall, accuracy, and F1-measure). First, results have already shown that some of these methods show higher prediction quality than the classical logit models.

CO110 Room 43 IASC JOURNAL OF DATA SCIENCE, STATISTICS, AND VISUALISATION (JDSSV) SESSION Chair: Stefan Van Aelst

C0232: **Robust multiway PCA for casewise and cellwise outliers**
Presenter: **Mehdi Hirari**, KU Leuven, Belgium
Co-authors: Stefan Van Aelst, Mia Hubert, Fabio Centofanti

Multi-way data extend two-way matrices to a higher-dimensional tensor. In many fields, it is relevant to pursue the analysis of such data by keeping it in its initial form without unfolding it into a matrix. Often, multi-way data are explored by means of dimensional reduction techniques. We study the Multilinear Principal Component Analysis (MPCA) model, which expresses the multi-way data in a more compact format by determining a multilinear projection that captures most of the original multi-way data variation. The most common algorithm to fit this model is an Alternating Singular Value Decomposition algorithm, which, despite its popularity, suffers from outliers. To address this issue, robust alternative methods were introduced to withstand casewise and cellwise outliers, respectively, where two different loss functions are tailored based on the type of outliers. However, such methods break when confronted with datasets contaminated by both types of outliers. To address this discrepancy, we propose a method by constructing a new loss function using M-estimators for multi-way data, offering robustness against both kinds of anomalies simultaneously. Extensive simulations show the efficacy of this Robust MPCA method against outliers, demonstrating its potential in robust multi-way data analysis.

C0248: **Comparison of interval time series**
Presenter: **Ann Maharaj**, Monash University, Australia
Co-authors: Paula Brito, Paulo Teles

An interval time series (ITS) consists of intervals observed at consecutive time points, with each interval defined by its lower and upper bounds or by its centre and radius. In practical scenarios, analysing an ITS can offer more valuable insights into the variability between upper and lower bounds at each time point compared to analysing traditional time series with single values at each time point. We compare two ITS by assessing the statistical significance of differences in their underlying distributions. To perform hypothesis testing, we use the discrete wavelet transform (DWT) which decomposes a time series into a set of coefficients over several frequency bands or scales. We perform randomisation tests on the DWT of the radius and centre of the two ITS at different scales. Randomisation tests require uncorrelated observations. This condition is more or less satisfied because at each scale, the DWT coefficients are approximately uncorrelated with each other. The proposed test statistic is the ratio of the determinants of the covariance matrix of radius and centre DWTs of the two ITS, at each scale. This test statistic ensures that the variability between the upper and lower bounds is captured. Through simulation studies to assess its performance, reasonably good estimates of size and power of the test are observed. Application of the test to real interval time series demonstrates its practical utility in analysing and comparing ITS effectively.

C0309: **Robust functional regression with discretely sampled predictors**
Presenter: **Ioannis Kalogridis**, KU Leuven, Belgium
Co-authors: Stanislav Nagy

The functional linear model is an important extension of the classical regression model allowing for scalar responses to be modeled as functions of stochastic processes. Yet, despite the usefulness and popularity of the functional linear model in recent years, most treatments, theoretical and practical alike, suffer either from (i) lack of resistance towards the many types of anomalies one may encounter with functional data or (ii) biases resulting from the use of discretely sampled functional data instead of completely observed data. To address these deficiencies, the aim is to introduce and study the first class of robust functional regression estimators for partially observed functional data. The proposed broad class of estimators is based on thin-plate splines with a novel computationally efficient quadratic penalty, is easily implementable and enjoys good theoretical properties under weak assumptions. We show that, in the incomplete data setting, both the sample size and discretization error of the processes determine the asymptotic rate of convergence of functional regression estimators and the latter cannot be ignored. These theoretical properties remain valid even with multi-dimensional random fields acting as predictors and random smoothing parameters. The effectiveness of the proposed class of estimators in practice is demonstrated by a simulation study and a real-data example.

C0324: **Is distance correlation robust?**
Presenter: **Jakob Raymaekers**, University of Antwerp, Belgium
Co-authors: Sarah Leyder, Peter Rousseeuw

Distance correlation is a popular measure of dependence between random variables. It has some robustness properties, but not all of them. We prove that the influence function of the usual distance correlation is bounded but that its breakdown value is zero. Moreover, it has an unbounded sensitivity function, converging to the bounded influence function for increasing sample size. To address this sensitivity to outliers, we construct a more robust version of distance correlation, which is based on a new data transformation. Simulations indicate that the resulting method is quite robust and has good power in the presence of outliers. We illustrate the method on genetic data. Comparing the classical distance correlation with its more robust version provides additional insight.

C0337: **Interpretable cost-sensitive ensembling**
Presenter: **Stefan Van Aelst**, University of Leuven, Belgium
Co-authors: Tim Verdonck, Bing Yang

In many applications, the cost of wrong decisions is not symmetric. Therefore, it makes sense to take the costs associated with wrong decisions into account in the decision process to minimize the risks for stakeholders. To achieve this, cost-sensitive methods have been developed, such as cost-sensitive logistic regression. Moreover, for high-dimensional data, ensemble models often yield a much better performance than a single (sparse) model. However, ensembles of a large number of models are difficult to interpret. A split-learning framework has been recently developed to combine the interpretability of a single model with the performance of ensemble models. This framework is used to introduce a diverse ensemble of cost-sensitive logistic regression models. This yields an ensemble that is interpretable with a low misclassification cost. To solve the non-convex optimization problem, a novel algorithm based on the partial conservative convex separable quadratic approximation is developed. The proposed method delivers outstanding savings, as demonstrated through extensive simulation and real-world applications in fraud detection and gene expression analysis.

CO086 Room 44 CLUSTERING OF COMPLEX DATA STRUCTURES Chair: Maria Brigida Ferraro

C0156: **Fuzzy clustering of circular time series based on a new dependence measure with applications to wind data**
Presenter: **Angel Lopez Oriona**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Time series clustering is an essential machine learning task with applications in many disciplines. While the majority of the methods focus on time series taking values on the real line, very few works consider time series defined on the unit circle, although the latter objects frequently arise in many applications. The problem of clustering circular time series is addressed. To this aim, a distance between circular series is introduced and used to construct a clustering procedure. The metric relies on a new measure of serial dependence considering circular arcs, thus taking advantage of the directional character inherent to the series range. Since the dynamics of the series may vary over time, we adopt a fuzzy approach, which enables the procedure to locate each series into several clusters with different membership degrees. The resulting clustering algorithm is able to group series generated from similar stochastic processes, reaching accurate results with series coming from a broad variety of models. A simulation study shows that the proposed method outperforms several alternative techniques besides being computationally efficient. An interesting application involving time series of wind direction in Saudi Arabia highlights the potential of the proposed approach.

C0312: **Biclustering of ordinal data through a composite likelihood approach**
Presenter: **Monia Ranalli**, Sapienza University of Rome, Italy
Co-authors: Francesca Martella

A finite mixture model to simultaneously cluster the rows and columns of a two-mode ordinal data matrix is proposed. Following the Underlying Response Variable (URV) approach, the observed variables are considered to be a discretization of latent continuous variables distributed as a mixture of Gaussians. To introduce a partition of the P variables within the g-th component of the mixture, we adopt a factorial representation of the data where a binary row stochastic matrix, representing variable membership, is used to cluster variables. In this way, we associate a component in the finite mixture with a cluster of variables and define a bicluster of units and variables. The number of clusters of variables (and therefore the partition of variables) may vary with clusters of units. Due to the numerical intractability of the likelihood function, the estimation of model parameters is based on composite likelihood (CL) methods. It essentially reduces to a computationally efficient Expectation-Maximization type algorithm. The performance of the proposed approach is discussed in both simulated and real datasets.

C0351: **Two-step clustering: A new method in the sequential deep clustering approach**
Presenter: **Claudia Rampichini**, University of Rome La Sapienza, Italy

The proposed method combines the use of two different clustering methods: fuzzy k-means and k-means. Membership degree values are used to identify units with an unclear assignment, and the crisp method is applied to this reduced dataset. A unit has an unclear assignment when membership degrees are close to each other. Knowing that one of the main problems of traditional clustering methods is related to the handling of high dimensional data, this proposal is combined with the use of a neural network according to the sequential deep clustering approach. Deep neural networks enhance clustering performance by reducing input data complexity, followed by clustering on the reduced dataset. An autoencoder neural network is utilized, and the two-step clustering method is applied to its results. As such, it is possible to obtain good results even with high-dimensional data such as images. Empirical studies demonstrate performance enhancements over individual k-means and fuzzy k-means methods, highlighting the effectiveness of neural networks in clustering tasks.

C0380: **Clustering three-way data**
Presenter: **Paul McNicholas**, McMaster University, Canada

Over the past few years, increased attention has been paid to the clustering of three-way data. Some such approaches are presented and discussed, with a focus on approaches based on (matrix-variate) mixture models.

C0381: **Unbiased mixed variables distance**
Presenter: **Michel van de Velden**, Erasmus University Rotterdam, Netherlands
Co-authors: Carlo Cavicchia, Alfonso Iodice D Enza, Angelos Markos

Defining a distance in a mixed data setting requires the quantification of observed differences of variables of different types and of variables that are measured on different scales. There exist several proposals for mixed variable distances, however, such distances tend to be biased towards specific variable types or measurement units. That is, the variable types or scales influence the contribution of individual variables to the overall distance. Unbiased mixed variable distance is defined as a distance for which the contributions of individual variables to the overall distance, are not influenced by measurement types or scales. The relevant concepts are defined to quantify such biases, and a general formulation is provided that can be used to construct unbiased mixed variable distances.

CO088 Room 45 RELIABLE PREDICTION MODELS FOR CHALLENGING DATA Chair: Garth Tarr

C0210: **Sparse-group SLOPE: Adaptive bi-level selection with FDR-control**
Presenter: **Fabio Feser**, Imperial College London, United Kingdom
Co-authors: Marina Evangelou

A new high-dimensional approach is proposed for simultaneous variable and group selection, called Sparse-group SLOPE (SGS). SGS achieves false discovery rate control at both variable and group levels by incorporating the sorted L-One penalized estimation (SLOPE) model into a sparse-group framework and exploiting grouping information. A proximal algorithm is implemented to fit SGS and work for both Gaussian and Binomial-distributed responses. Penalty sequences specific to SGS were derived and shown to provide FDR control under orthogonal designs. Through the analysis of both synthetic and real datasets, the proposed SGS approach is found to outperform other existing lasso- and SLOPE-based models for bi-level selection and prediction accuracy. Further, the problem of model selection is investigated with regard to FDR-control through the choice of the tuning parameter. Various model selection and noise estimation approaches for selecting the tuning parameter of the regularisation model are proposed and compared in a simulation study. Additionally, a new adaptive noise estimation procedure is proposed for SGS, termed Adaptively Scaled SGS (AS-SGS), and is an extension of the scaled lasso.

C0212: **Subset selection via continuous optimization**
Presenter: **Samuel Muller**, Macquarie University, Australia
Co-authors: Benoit Liquet, Sarat Moka, Houying Zhu

Recent rapid developments in information technology have enabled the collection of high-dimensional complex data, including in engineering, economics, finance, biology, and health sciences. High-dimensional means that the number of features is large and often far higher than the number of collected data samples. In many of these applications, it is desirable to find a small best subset of predictors so that the resulting model has desirable prediction accuracy. We will first briefly review existing optimization and search methods in the literature that tackle the problem of identifying or selecting the set of important predictors. We then present the COMBSS framework, a continuous optimization-based solution that we recently showed to solve the best subset selection problem in linear regression. Then, we focus on highlighting how COMBSS can be extended to other models. We explore how this is possible in generalized linear models, partial least-squares or principal component analysis.

C0221: **Computational strategies for regression model selection in the high-dimensional case**
Presenter: **Marios Demosthenous**, Justus Liebig University of Giessen, Germany
Co-authors: Cristian Gatu, Erricos Kontoghiorghes

Computational strategies for finding the best-subset regression models are proposed. The case of high-dimensional (HD) data where the number of variables exceeds the number of observations is considered. Within this context, a theoretical combinatorial solution is proposed. It is based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is adapted to the HD case. Furthermore, the R package lmSubsets is employed in the HD case to identify the best submodel based on the AIC family selection criteria. Preliminary experimental results are presented and analyzed. The efficient extension of the lmSelect algorithm to HD is discussed.

C0273: **Efficient stability screening for ultra-high dimensional data**
Presenter: **Ibrahim Joudah**, Macquarie University, Australia
Co-authors: Samuel Muller, Houying Zhu

Data is ever more complex and high-dimensional in many fields, such as genomics, social science, health, and finance. This presents exciting challenges for statistical analysis. Stability selection, a technique used to stably identify important features, struggles with high dimensionality, often already when the number of features is in their thousands, but even more so when it well exceeds tens of thousands. To address this, stability screening is proposed to screen features stably and efficiently prior to implementing variable selection. Stability screening is a feature screening approach that relies on efficient subsampling techniques, aiming to facilitate stable selection after the initial screening. The latest findings from ongoing research into feasible stability screening are presented. The proposed method for stability screening is illustrated using both simulated and real-world data.

C0373: **Assessment of case influence in the Lasso with a case-weight adjusted solution path**
Presenter: **Zhenbang Jiao**, The Ohio State University, United States
Co-authors: Yoonkyung Lee

Case influence in the Lasso regression is studied using Cook's distance, which measures the overall change in the fitted values when one observation is deleted. Unlike in ordinary least squares regression, the estimated coefficients in the Lasso do not have a closed form due to the nondifferentiability of the l1 penalty, and neither does Cook's distance. To find the case-deleted Lasso solution without refitting the model, a weight parameter ranging from 1 to 0 is introduced to approach it from the full data solution and generate a solution path indexed by this parameter. It is shown that the solution path is piecewise linear with respect to a simple function of the weight parameter under a fixed penalty. The resulting case influence is a function of the penalty and weight parameters, and it becomes Cook's distance when the weight is 0. As the penalty parameter changes, selected variables change, and the magnitude of Cook's distance for the same data point may vary with the subset of variables selected. In addition, a case influence graph is introduced to visualize how the contribution of each data point changes with the penalty parameter. From the graph, influential points can be identified at different levels of penalization and make modeling decisions accordingly. Moreover, it is found that case influence graphs exhibit different patterns between underfitting and overfitting phases, which can provide additional information for model selection.

CC055 Room 050 HIGH-DIMENSIONAL STATISTICS Chair: Stefan Sperlich

C0241: A unified class of null proportion estimators with plug-in FDRcontrol

Presenter: **Sebastian Doehler**, Darmstadt University of Applied Science, Germany

The Benjamini-Hochberg (BH) procedure is a staple of modern high-dimensional data analysis. This method can be made more powerful by incorporating estimators of the proportion of null hypotheses, yielding an adaptive BH procedure which still controls the false discovery rate (FDR). We present a unified class of estimators, which comes with mathematical guarantees, encompasses existing and new estimators and can also be adapted to discrete tests. While our focus is on presenting the generality and flexibility of this new class, we also include someanalyses on simulated and real data.

C0478: Noise and overfitting: A new perspective on the predictive performance of a linear model

Presenter: **Insha Ullah**, Australian National University, Australia

Co-authors: Alan Welsh

Traditionally, the bias-variance trade-off has guided model selection in under-parameterized regimes, with the belief that overparameterization leads to overfitting and poor generalization. However, recent studies have uncovered the double descent curve, where test error surprisingly decreases in overparameterized models, challenging this framework. The aim is to examine the counterintuitive benefits of overfitting in linear models, investigating how noise from predictors or observations affects prediction accuracy. This exploration is crucial, as irrelevant variables are common in practical applications yet often overlooked in discussions on the double descent curve and regularization techniques like ridge regularization. The findings explain the double descent curve mechanics and suggest that overfitting can enhance prediction accuracy under certain conditions. Recent research has shown that minimum norm least squares estimation performs shrinkage in the presence of irrelevant predictors and tends to outperform ridge regularization with a positive ridge penalty in terms of prediction accuracy. Empirical evidence also suggests that the optimal ridge penalty may be zero or negative, challenging standard practice. The analysis demonstrates the advantages of a negative ridge penalty, highlighting the role of noise in model performance.

C0435: Graph-linked unified embedding considering label information

Presenter: **Hiroshi Kobayashi**, Doshisha University, Japan

Co-authors: Masaaki Okabe, Hiroshi Yadohisa

Graph-Linked Unified Embedding (GLUE) estimates the low-dimensional space shared across datasets derived from multiple sources. GLUE leverages prior knowledge between variables, represented as a knowledge graph and employs graph neural networks to achieve dimension reduction while considering latent structures. This method is particularly effective for analyzing related datasets and is often applied to multi-omics data, which integrates various omics data such as gene expression, proteomics, and DNA methylation. Unlike single-omics analysis, multi-omics data collected through diverse experiments or single-cell measurements contain complementary information that can improve the understanding of complex biological systems and diseases. However, applying GLUE to datasets with staging or cell-type labels may lead to low-dimensional representation that overlooks the label information. To address this, conditional GLUE (CGLUE) is proposed. By employing a conditional variational autoencoder, CGLUE conditions the encoder with labels and integrates multiple datasets into a shared low-dimensional space across sources while preserving label information. This approach promotes the close positioning of data points with the same label in a low-dimensional space, enhancing the interpretability of data based on label-specific information through visualization and analysis.

C0502: Optimizing interval PLS via GP regression

Presenter: **Nicolas Hernandez**, Queen Mary University of London, United Kingdom

Co-authors: Tom Fearn, Yoonsun Choi

Interval partial least-squares regression (iPLS) is an adaptation of the partial least squares regression (PLS) tailored for high dimensional spectral data, such as near-infrared spectra. Spectrometric data is expressed over a continuous domain. Therefore, interval selection is a more viable alternative for feature extraction than variable selection. Despite its potential, a primary challenge in iPLS remains in the selection of optimal intervals. Although traditional approaches, such as forward and backward selection methods, have practical benefits, they have crucial limitations of heavy reliance on heuristic approaches. The aim is to propose a novel approach to interval selection in iPLS via uncertainty quantification techniques. Gaussian process regression is used, emphasizing its ability for flexible modelling and its provision of uncertainty estimates. This integration aims to optimize the accuracy of interval selection to highlight discrepancies between model predictions and observations. The contribution is in evolving dialogue on improving spectral data analysis techniques in the iPLS domain, with an application to the Spectrometric field.

C0186: Accounting for population heterogeneity by modeling interactions with the pliable lasso

Presenter: **Theophilus Quachie Asenso**, University of Oslo, Norway

Co-authors: Manuela Zucknick

The pliable lasso penalty is applied to estimate interaction effects and extend the existing linear pliable lasso model to the multi-response problem. In the first part, results from the recent work on the regularized multi-response regression problem are presented, where there exists some structural relation within the responses and between the covariates and a set of modifying variables. To handle this problem, MADMMplasso is proposed, a novel regularized regression method. This method is able to find covariates and their corresponding interactions, with some joint association with multiple related responses. The interaction term is allowed between the covariate and modifying variable to be included in a weak asymmetrical hierarchical manner by first considering whether the corresponding covariate main term is in the model. The results from the simulations and analysis of a pharmacogenomic screen data set show that the proposed method has an advantage in handling correlated responses and interaction effects, both with respect to prediction and variable selection performance. In the second part, results are reported from ongoing work from the implementation of the MADMMplasso in modelling and predicting synergistic effects between two drugs in drug combination experiments, using, for example, the molecular characterization of a cell line with multi-omics data to predict whether two drugs will act synergistically on that particular cell line.

CC135 Room 051 MACHINE LEARNING FOR APPLICATIONS Chair: Florian Frommlet

C0191: Impact of rarity level and resampling techniques on machine learning classification performance

Presenter: **Olcay Alpay**, Sinop University, Turkey

The classification of binary events is frequently discussed in the literature. However, in cases where the distribution of the event of interest is unbalanced, such as with rare events, machine learning algorithms may produce biased results. The classification performance of several machine learning algorithms is investigated, including Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, and Artificial Neural Network, in different rarity scenarios and how they are affected by unbalanced data. The impact of resampling techniques on dataset balancing and classification is also examined.

C0194: **On some mathematical foundations of machine learning algorithm and an application**
Presenter: **Ismail Aydin**, Sinop University, Turkey
Co-authors: Olcay Alpay

Machine learning is an interdisciplinary research field that includes probability and statistics, linear algebra, calculus (gradient descent), nonlinear partial differential equations (stochastic problems), Fourier transform, signal processing and computer science, among others. The application of machine learning/deep learning algorithms and methods has become widespread in areas such as mental health diagnosis, object recognition, image processing, semantic segmentation, human action recognition, finance, social sciences, operations research, and epidemic management. Machine learning is based on three concepts: data, models, and learning. As in all everyday applications, the amount of data in scientific experiments has increased significantly compared to a decade ago. Although it is expensive and time-consuming to analyze this large amount of data, machine learning algorithms can provide efficient and fast results. We present some mathematical foundations and methods for machine learning algorithms and an application to a suitable dataset.

C0454: **Correlates of suicide ideation among young adults: Insights from machine learning algorithms**
Presenter: **Mogana Darshini Ganggayah**, Monash University Malaysia, Malaysia
Co-authors: Erniel Barrios, Hariharan Muniandy

Target 3.4 (including suicide mortality rate as one indicator) of sustainable development goals set to reduce premature death from noncommunicable diseases by one-third in 2030. In 2019, global suicide rates were estimated at 9.2 per 100,000 population; this is highest in Europe at 12.8 and also relatively higher in Southeast Asia at 10.1. Among adolescents (15-24 years old), the suicide rate is 7.97 and exceeds over 20 in many countries. Suicide was the fourth leading cause of death among 15-29 years old in 2016. Many suicides happen impulsively in moments of crisis, which is triggered by poor mental health. Some machine learning algorithms are used to identify possible risk factors associated with suicide ideation among Filipino youths based on the 5th Young Adult Fertility and Sexuality Survey (YAFS5). Some indicators related to personal experience, internet usage (especially heavy engagement in social media), intake of alcohol, marital status, experience of sexual harassment, and sleep difficulties lead to distress associated with mental health conditions that are associated with suicidal ideation. This underscores the necessity for a holistic approach to suicide prevention that addresses a wide spectrum of risk factors. The significant role of statistical machine learning is also exhibited in further extracting insights from survey data.

C0286: **Deep learning applications in mental workload classification**
Presenter: **Serenay Cakar**, Middle East Technical University, Turkey
Co-authors: Fulya Gokalp Yavuz

The $n-$back task paradigm provides rich temporal data, capturing nuanced insights into working memory and cognitive demand across different conditions. Introducing innovative perspectives, deep learning techniques are integrated to classify mental workload levels from n-back cognitive data with dense and sparse features. Our findings highlight the efficacy of the extreme Deep Factorization Machine (xDeepFM) model, validated through stratified 5-fold cross-validation. Compared to the baseline model for the 0- vs 1-back classification task, this approach yielded substantial enhancements: 67.50% accuracy, 68.74% sensitivity, 66.24% specificity, and 68.48% F1-Score. Notably, dense features comprising combinations of hemodynamic measures and experimental variables, with subjects as sparse features, contributed significantly to these improvements. Additionally, incorporating Principal Component Analysis (PCA) led to notable enhancements: 53.03% accuracy, 98.03% sensitivity, and 70.37% F1-Score, particularly evident in classifying the 0- vs 1-back condition using the xDeepFM model. These outcomes underscore the significant role of deep learning methodologies in accurately classifying mental workload levels from complex $n-$back cognitive science data, providing invaluable insights into cognitive functioning and workload assessment.

C0376: **Weighted robust hybrid partial least squares regression forest**
Presenter: **Aylin Alin**, Dokuz Eylul University, Turkey

Partial least squares regression(PLSR) is a widely utilized technique for modeling data characterized by multicollinearity or instances where the predictor count exceeds the number of observations. In parallel, random forest regression or regression forest (RF) is an ensemble method proficient in managing extensive datasets, addressing missing predictor values, and mitigating multicollinearity issues. The integration of PLSR and RF termed hybrid PLSR-RF amplifies modeling efficacy. Notwithstanding their advantages, PLSR, RF, and the hybrid PLSR-RF remain susceptible to outlier influence. Although robust variants of PLSR and RF exist, the literature lacks robustified iterations of the hybrid approach. The robust hybrid partial least squares regression forest methodology is introduced to address this gap. This novel method leverages the robust iteratively reweighted SIMPLS algorithm (RWSIMPLS) to derive orthogonal components that subsequently inform the construction of a regression forest. Weighted predictions are applied within this forest to diminish outlier impact in individual trees. Moreover, an alternative bagging technique is proposed that mitigates outlier effects and constrains tree complexity.

Wednesday 28.08.2024 09:00 - 10:30 Parallel Session E – COMPSTAT2024

CI009 Room 45 LEARNING FROM MACHINE LEARNING BY STRUCTURING Chair: Stefan Sperlich

C0168: **Local machine learning for data giants**
Presenter: **Michael Scholz**, University of Klagenfurt, Austria
Co-authors: Stefan Sperlich, Gilles Cattani

Classical nonparametric estimation is the natural link between Breiman's two cultures, say 'traditional regressions methods' and 'pure prediction algorithms'. We borrow ideas of local smoothers and efficient implementation to combine good practices of both cultures for generating a practical tool for the statistical analysis of large data problems, may it be estimation, prediction or attribution. Estimation and prediction are particularly successful when allowing for local adaptiveness. Further, while typically distributed databases are considered a bane, data localization can turn it into a boon. Similarly, since most of the problems with divide-and-conquer algorithms are rooted in the paradigm of facing a global parameter set, they disappear by localization, and the selection of an optimal subsample size is melted with the one of optimal bandwidths which, in addition, we allow to be local too. Moreover, model and variable selection are possible, and sometimes even necessary, when staying local. For each step and subprocedure, we look for the most efficient implementation to keep the procedure fast. The proof of concept and computational details are given in a simulation study. An application to ocean warming illustrates the practical use of such a tool.

C0223: **Random planted forest**
Presenter: **Munir Hiabu**, University of Copenhagen, Denmark
Co-authors: Joseph Meyer, Enno Mammen

A novel interpretable tree-based algorithm is introduced for prediction in a regression setting. The motivation is to estimate the unknown regression function from a functional decomposition perspective in which the functional components correspond to lower-order interaction terms. The idea is to modify the random forest algorithm by keeping certain leaves after they are split instead of deleting them. This leads to non-binary trees, which we refer to as planted trees. An extension to a forest leads to our random planted forest algorithm. Additionally, the maximum number of covariates which can interact within a leaf can be bounded. If we set this interaction bound to one, the resulting estimator is a sum of one-dimensional functions. In the other extreme case, if we do not set a limit, the resulting estimator and corresponding model place no restrictions on the form of the regression function. In a simulation study, we found encouraging prediction and visualisation properties in our random planted forest method. We also develop theory for an idealized version of random planted forests in cases where the interaction bound is low. We show that if it is smaller than three, the idealized version achieves asymptotically optimal convergence rates up to a logarithmic factor. The code is available on GitHub.

C0244: **A complete guide to small area learning**
Presenter: **Katarzyna Reluga**, University of Bristol, United Kingdom

Sample surveys are widely recognized as high-quality and cost-effective sources of information for obtaining estimates of target parameters at the population and subpopulation levels. If the sample size of the subpopulation is small (or even zero in some areas), researchers encounter the small area estimation (SAE) dilemma. SAE techniques have been developed to provide official statistics by leveraging survey samples and parametric statistical modelling. We introduce a general framework for small area learning (SAL). SAL encompasses machine learning to obtain estimates of subpopulation-level parameters by pooling information from other subpopulations, which is the main principle of classical SAE. In addition to presenting a complete methodological setup for inference and prediction, we provide a practical application of SAL in measuring poverty.

CO087 Room 052 RELIABLE AND ACCURATE STATISTICAL SOLUTIONS FOR MODERN COMPLEX DATA Chair: Samuel Muller

C0267: **Continuous optimization for offline change point detection and estimation**
Presenter: **Hans Reimann**, University of Potsdam, Germany
Co-authors: Sarat Moka, Georgy Sofronov

The application of novel advances in best subset selection for regression modelling is explored via continuous optimization for offline change point detection and estimation in univariate Gaussian data sequences. The main idea hereby lies in reformulating the normal mean multiple change-point model into a regularized statistical inverse problem enforcing the sparsity of the parameter vector. After introducing the problem statement, criteria and recalling previous investigation via Lasso-regularization for sparsity, the novel and enabling framework of continuous optimization for best subset selection (COMBSS) is briefly introduced and connected to the problem at hand. Both supervised and unsupervised perspectives are explored, with the latter testing different approaches for the choice of regularization penalty parameters via the discrepancy principle and a confidence bound. The main result is an adaptation and evaluation of the COMBSS approach for offline normal mean multiple change-point detection via experimental results on simulated data for different choices of regularisation parameters. Results, as well as further directions for investigations, are then critically discussed.

C0277: **Visualising model selection stability**
Presenter: **Garth Tarr**, University of Sydney, Australia

Various proposals to understand the stability of selected features from various models have been proposed over the past 15 years. Some of these are investigated from the lens of how to communicate selection stability results in a visual way to aid understanding and interpretation for statistical practitioners. A key tool is the mplot R package which has the aim of helping researchers implement stability selection concepts to use to better inform the variable selection process. The focus is on recent improvements to the mplot package that builds on recent advances in exhaustive search techniques and on separating out stably selected features in regularized regression models.

C0290: **Misspecification matters: Prediction under misspecified random effects distributions in GLMMs**
Presenter: **Quan Vu**, Australian National University, Australia
Co-authors: Francis Hui, Samuel Muller, Alan Welsh

The generalized linear mixed model (GLMM) is widely used in applied sciences because of its capability to model clustered data. One important aspect when dealing with GLMMs is the prediction of random effects and mean responses. There have been contradictory views in the literature on whether the normality assumption on the random effects significantly impacts the quality of the prediction with respect to mean squared prediction error (MSEP) when the underlying random effects are not normal. We investigate this problem by comparing the empirical best predictors of the random effects and the mean responses under a misspecified normal distribution against those under a correctly specified distribution, which is a mixture of normal distributions. Our findings indicate that the unconditional MSEPs for the random effects are higher under the incorrectly assumed normal distribution, when the true random effects distribution is very skewed or multimodal, especially when the cluster size is small. The conditional MSEPs for the random effects are also generally higher under the misspecified distribution, especially at the region closer to the mean of each component of the underlying mixture distribution (given this distribution is skewed or multimodal). These results demonstrate the importance of random effects specification to prediction in GLMMs.

C0219: **A multiplicative semiparametric regression solution for non-Euclidean data**
Presenter: **Luca Maestrini**, The Australian National University, Australia
Co-authors: Janice Scealy, Francis Hui, Andrew Wood

In many regression applications involving non-Euclidean response variables, it is important to have available models that have sufficient flexibility to accommodate both local and global features. In models for local features, the regression function is assumed to be a general unknown function defined on the non-Euclidean geometric space, which can be estimated using a smoothing method. In global models, a parametric form is specified for the regression function, for example by using a known link function mapping linear combinations of regression coefficients and covariates onto the non-Euclidean space. Existing models are either entirely global or entirely local, and to overcome this problem, we develop local-global regression models for non-Euclidean response variables following an extrinsic approach, i.e. using an ambient space metric. For non-Euclidean spaces with sufficiently rich isometry groups, such as spheres, it is possible to separate the non-parametric and parametric components in the regression function via multiplicative models. We exploit this multiplicative structure to make our formulation more computationally advantageous. Non-linear least squares can be used to estimate the unknown parameters in the parametric part, and the nonparametric part can be estimated in the Euclidean space using penalised splines and fitted using standard linear mixed effects model software.

CO091 Room 43 STATISTICAL MODELS AND ALGORITHMS FOR SURVIVAL DATA Chair: Ambra Macis

C0238: **Feature screening and selection in competing risks models**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy

In the analysis of time-to-event data, competing risks data are encountered when individuals may fail from multiple causes (for example, $K > 2$), and the occurrence of one failure event precludes the others from happening. Different (un)correlated features should influence the events, and the same feature should affect more than one event. Moreover, the number of covariates (p) should be very large and sometimes should be greater than the sample size (n). Thus, a reduction of variables is crucial. In the literature, some authors focused on screening and selecting the variables under the assumption that a) the number of events K is 2, and b) one event is of the main interest, while the other can be neglected. Thus, the aims are to i) compare the performance of some existing methods for screening and selecting the most significant variables, ii) highlight their main advantages and disadvantages, and iii) propose a new procedure able to identify the relevant covariates in the framework of high and ultra-high dimensions, in the presence of highly correlated variables, and when the number of events K is larger than two.

C0272: **Estimation and log-rank testing procedure via bivariate survival copula models under semi-competing risk**
Presenter: **Tomoyuki Sugimoto**, Osaka University, Japan

In time-to-event data, two primary endpoints of interest are often non-fatal and fatal, and then the issue of semi-competing risks arises when designing trials or performing statistical analyses. If such non-fatal and fatal event-times are mutually uncorrelated, we can conduct the usual statistical analysis. However, when the two event times are correlated, the usual use of log-rank analysis or Cox regression models for non-fatal events will reduce the power of the analysis. The problem of estimating and testing the hazard ratio of the marginal distribution of the non-fatal event is difficult to handle not only in theory but also in practical use. Assuming a bivariate survival copula model in which two event times are possibly correlated under semi-competing risks, we propose an estimation method and algorithm for making an inference on the marginal distribution of the non-fatal events. We discuss the theoretical validity and theoretical properties of the estimated survival function based on the proposed estimation procedure. Furthermore, we discuss the extension to the two-sample problem via the estimation for survival functions and copula-related parameters. We discuss a modified version of the log-rank analysis and evaluate the performance of this estimation procedure using bivariate survival copulas.

C0336: **Boosting distributional copula regression for right-censored bivariate time-to-event data**
Presenter: **Guillermo Briseno Sanchez**, TU Dortmund University, Germany
Co-authors: Nadja Klein, Andreas Mayr, Andreas Groll

A highly flexible distributional copula regression model is proposed for bivariate right-censored time-to-event data. The joint survival function is constructed using parametric copulas, allowing for separate specifications of the dependence structure between the time-to-event variables and their respective marginal survival distributions. The latter can be specified using well-known parametric distributions, such as log-normal, log-logistic, or Weibull distributions. These results were then in parametric (accelerated failure time, AFT) models for the respective univariate responses. By embedding the model into the framework of generalised additive models for location, scale, and shape (GAMLSS), all parameters of the joint distribution can be modeled as a function of covariates. Using additive predictors thereby enables the account of linear, non-linear, or spatial effects in modelling the dependence structure and the respective marginal distributions. Estimation is proposed by means of component-wise gradient-based boosting to allow for data-driven variable selection. The latter not only renders model building feasible and avoids the manual comparison of different model specifications but also allows the tackling of high-dimensional (p » n) data structures. To the best of knowledge, this is the first implementation of multivariate AFT models via distributional copula regression and automatic variable selection via statistical boosting.

C0365: **Predicting patient trajectories with deep multi-state models trained on electronic health record data**
Presenter: **Thomas Matcham**, Imperial College London, United Kingdom

In recent years, a range of survival models incorporating deep learning elements have been produced to better model survival data, particularly when very large quantities of data are available. Electronic health records contain enormous quantities of patient trajectory data, giving time-to-event data for the progression of diseases as well as consecutive interactions with healthcare providers. The extension of deep learning survival models is explored in the multi-state setting, modelling long-term COVID-19 hospitalization outcomes and the progression of type II diabetes-related chronic diseases.

CO090 Room 44 ADVANCES IN FUNCTIONAL STATISTICS Chair: Enea Bongiorno

C0296: **Fourier approach to goodness-of-fit tests for Gaussian random processes**
Presenter: **Daniel Hlubinka**, Univerzita Karlova, Czech Republic
Co-authors: Zdenek Hlavka, Petr Coupek, Viktor Dolnik

A new goodness-of-fit (GoF) test is proposed and investigated for the Gaussianity of the observed functional data. The test statistic is the Cramr-von Mises distance between the observed empirical characteristic functional (CF) and the theoretical CFcorresponding to the null hypothesis stating that the functional observations (process paths) were generated from a specific parametric family of Gaussian processes, possibly with unknown parameters. The asymptotic null distribution of the proposed test statistic is also derived in the presence of these nuisance parameters, the consistency of the classical parametric bootstrap is established, and some particular choices of the necessary tuning parameters are discussed. The empirical level and power are investigated in a simulation study involving GoF tests of an Ornstein-Uhlenbeck process, Vasicek model, or a (fractional) Brownian motion, both with and without nuisance parameters, with suitable Gaussian and non-Gaussian alternatives.

C0333: **Spherical functional autoregressive models for global aircraft-based atmospheric measurements**
Presenter: **Alessia Caponera**, LUISS Guido Carli, Italy
Co-authors: Almond Stoecker, Victor Panaretos

Motivated by global analysis of aircraft-based measurements of air pollutants and climate variables, specifically the COVID-19 pandemic's possible impact on ozone concentrations, a functional autoregressive model is proposed to capture global spatiotemporal variability, incorporating solar radiation cycles. Efficient estimation techniques are developed, and means of suitable visualization are demonstrated, paving the way for similar analyses in the future.

C0342: **fsemipar: An R package for SoF semiparametric regression**
Presenter: **Silvia Novo**, Universidade da Coruna, Spain
Co-authors: German Aneiros

Functional data analysis has become a tool of interest in applied areas such as economics, medicine, and chemistry. Among the techniques developed in recent literature, functional semiparametric regression stands out for its balance between flexible modelling and output interpretation. Despite the large variety of research papers dealing with scalar-on-function (SoF) semiparametric models, there is a notable gap in software tools for their implementation. The R package fsemipar, tailored for these models, is introduced. fsemipar not only estimates functional single-index models using kernel smoothing techniques but also estimates and selects relevant scalar variables in semi-functional models with multivariate linear components. A standout feature is its ability to identify impact points of a curve on the response, even in models with multiple functional covariates, and to integrate both continuous and pointwise effects of functional predictors within a single model. In addition, it allows the use of location-adaptive estimators based on the k-nearest-neighbors approach for all the semiparametric models included. Its flexible interface empowers users to customize a wide range of input parameters and includes the standard S3 methods for prediction, statistical analysis, and estimate visualization (predict, summary, print, and plot), enhancing clear result interpretation.

C0360: **Transportation-based change point detection and testing for functional covariances**
Presenter: **Valentina Masarotto**, Leiden University, Netherlands

The analysis of variation within a sample of stochastic processes is addressed, particularly focusing on their second-order covariance structure. Covariance operators are primarily known in functional data analysis for the crucial role they play in the Karhunen-Love expansion. However, these operators themselves may exhibit variability and necessitate statistical techniques tailored to assess their fluctuations. Such techniques are closely tied to the choice of metric on covariance operators and, if mastered, grant access to powerful statistical procedures. The geometric properties of the space of functional covariances are leveraged to develop inferential tools for such operators. In particular, by identifying covariances with centered Gaussian processes, results from optimal transport theory are exploited, in addition to functional data analysis. A novel approach is introduced to k-sample testing in the functional setting, which, in turn, will be applied to the detection of structural breaks in the covariance of a sample of functional data. By navigating the complex relationship between geometry, statistical theory, and functional analysis, the aim is to provide a systematic framework for nuanced inference and robust detection within the realm of stochastic processes. All algorithms presented are illustrated using real data and are implemented in the R package fdWasserstein.

CC011 Room 001 TIME SERIES Chair: Roland Fried

C0169: **Some contributions to harmonizable time series analysis**
Presenter: **Jean-Marc Freyermuth**, Aix-Marseille University, France
Co-authors: Anna Dudek, Dominique Dehay

Harmonizable time series are natural extensions of stationary time series with a spectral decomposition whose components are correlated. Thus, the covariance function of a harmonizable time series is bivariate and admits a two-dimensional Fourier decomposition (Loeve spectrum). They form a broad class of nonstationary processes that has been a subject of investigation for a long time. We introduce a parametric form for these harmonizable processes, namely Harmonizable Vector AutoRegressive and Moving Average models (HVARMA), and we give tools to generate finite time sample realizations of HVARMA with known Loeve spectrum. Then, we discuss nonparametric estimation of spectral characteristics of spatiotemporal processes that are locally time-harmonizable, and illustrate its application in EEG data analysis.

C0206: **Quasi-maximum likelihood estimation of causal linear long memory processes**
Presenter: **Jean Marc Bardet**, University Paris Pantheon-Sorbonne, France
Co-authors: Yves Gael Tchabo MBienkeu

The purpose is to study the convergence of the quasi-maximum likelihood (QML) estimator for long-memory linear processes. We first establish a correspondence between the long-memory linear process representation and the long-memory AR(∞) process representation. We then establish the almost sure consistency and asymptotic normality of the QML estimator. Numerical simulations illustrate the theoretical results and confirm the good performance of the estimator.

C0416: **Singular vector autoregressions**
Presenter: **Rodney Strachan**, The University of Queensland, Australia
Co-authors: Eric Eisenstat

The purpose is to develop methods for the empirical analysis of singular processes. A strong rationale, a well-developed theoretical framework, and, as shown, empirical support exist for multivariate time series with a singular spectral density. A singular spectral density is consistent with the economic theory underlying, for example, DSGE models, in which the number of variables is greater than the number of structural shocks. This assumption guarantees the existence of a finite order VAR representation, but a unique probability density function does not exist with respect to the Lebesgue measure. A density on a compact submanifold is therefore defined with respect to the Hausdorff measure, and in a Bayesian framework, an HMC algorithm is developed that jointly samples coefficients, lag length, and the number of shocks. The proposed framework is used to carry out a structural analysis of the US macroeconomy with COVID-19 shocks.

C0402: **Outlier-robust estimation of state-space models using a penalized approach**
Presenter: **Rajan Shankar**, University of Sydney, Australia
Co-authors: Garth Tarr, Ines Wilms, Jakob Raymaekers

State-space models are a broad class of statistical models for time-varying data. The Gaussian distributional assumption on the disturbances in the model leads to poor parameter estimates in the presence of additive outliers. Whilst there are ways to mitigate the influence of outliers via traditional robust estimation methods such as M-estimation, this issue is approached from a more modern perspective that utilizes penalization. A shift parameter is introduced at each timepoint, with the goal being that outliers receive a non-zero shift parameter while clean timepoints receive a zero shift parameter after estimation. The vector of shift parameters is penalized to ensure that not all shift parameters are trivially non-zero. Apart from making it feasible to fit accurate and reliable time series models in the presence of additive outliers, other benefits of this approach include automatic outlier flagging and visual diagnostic tools such as BIC curves to provide researchers and practitioners with better insights into the outlier structure of the data.

CC140 Room 050 CATEGORICAL DATA ANALYSIS Chair: Claudia Kirch

C0457: **Bayesian mixture SEM for ordinal categorical data**
Presenter: **Hiroki Takeshima**, Doshisha University Graduate School, Japan
Co-authors: Jun Tsuchida, Hiroshi Yadohisa

To compare the relationships of constructs between different groups, data from each group is collected using a common questionnaire, such as five-point Likert scale questions, and applied multi-group structural equation modeling (MGSEM). However, when the sample size of each group is small, the parameter estimation of the MGSEM tends to be unstable. In addition, the parameter estimation performance of the MGSEM tends to decrease when it is applied to ordinal categorical data. To address these issues, in this report, an MGSEM is proposed for ordinal categorical data with clustering of groups. Specifically, a mixture model is used to represent the differences in the relationships of constructs between groups. The mixture model allows the incorporation of information on parameters from groups within the same cluster. This is expected to stabilize the parameter estimation, even for groups with small sample sizes. To treat ordinal categorical variables as continuous, a continuous latent variable behind each ordinal categorical variable is assumed. Using these continuous latent variables for continuous SEM is expected to improve the estimation performance. The results of numerical experiments show that the parameter estimation performance of the proposed method is superior to that of the MGSEM.

C0471: **The hierarchical clustering-based method powered by the bootstrap approach for multiple imputations in categorical data**
Presenter: **Jaroslav Hornicek**, Prague University of Economics and Business, Czech Republic
Co-authors: Zdenek Sulc, Hana Rezankova, Jana Cibulkova

The imputation of missing values in nominal variables is a crucial yet underexplored area of research. A novel imputation method based on agglomerative clustering was proposed. This method clusters objects in the dataset using modified techniques to evaluate the similarity between objects with missing values, followed by imputing the missing values based on the derived hierarchical scheme. To enhance this approach, a bootstrap method was incorporated to enable multiple imputation. After that, two sets of simulated data were generated: one with missing values under the missing not at random mechanism and another under the missing at random mechanism. The imputation results on these sets were compared with those obtained using multiple imputations by the MICE and the EM algorithms, applying various evaluation criteria. The proposed techniques were implemented using advanced programming tools to increase computational speed, such as the Rcpp package, which integrates C++ within the R environment. The novel approach performs comparably to established algorithms, offering the additional advantage of being nonparametric. The results also showed the significant influence of the ratio of missing values, the number of categories in variables, and the moderate impact of the strength of association between variables.

C0464: **A simplification of aggregated symbolic data**
Presenter: **Junji Nakano**, Chuo University, Japan
Co-authors: Nobuo Shimizu, Yoshikazu Yamamoto

The interest is in comparing groups of individuals, where each individual is described by observations of continuous and categorical variables. To summarize each group, the number of individuals, the first and second moments of continuous variables and dummy variables for categorical variables are used. Such statistics are called aggregated symbolic data (ASD). Although ASD is an appropriate summary of a group, it is still complicated. There is a need to simplify it more for intuitive understanding and visualization. The aim is to treat continuous variables and categorical variables equally in the simplification by defining appropriate scores for categorical values. The method of multiple correspondence analysis is used to determine scores for categorical values. A visualization of the simplified ASD is also presented.

CC134 Room 051 APPLIED AND EMPIRICAL STATISTICS Chair: Andreas Artemiou

C0193: **Performance evaluation of some modified maximum likelihood estimators for power function distribution with outliers**
Presenter: **Demet Han Aydin**, Sinop University, Turkey

The estimation procedure for the parameters of the probability distribution is essential for statistical inference, especially for testing the hypotheses and establishing confidence intervals. One of the widely-used statistical distributions in the context of reliability studies is the Power Function distribution because it allows flexible modelling of lifetime data. In the presence of outliers, the performances of modified maximum likelihood estimators of the parameters of the power function distribution are studied. A Monte-Carlo simulation study is designed and conducted to evaluate the performances of the estimators of the parameters. Next, the wind dataset modified by outliers is analysed to corroborate the simulation results and to demonstrate the usefulness of the estimators.

C0431: **Employment of tertiary education graduates: International statistical comparisons**
Presenter: **Maria Frolova**, SquirrLE school, Thailand

Statistical patterns are identified and evaluated in the relationship between the level of economic development of a country and the share of the population with tertiary education diplomas, as well as the composition of university graduates by the education levels, using the data of OECD countries. Cross-country comparative elasticity of employment is analyzed by the level and composition of the working-age population (tertiary education graduates). The choice of educational program level depends on the current economic situation in the country, certain demand for labor force, the income level of households, as well as the ability of the population to foresee the leading industries and prospects for the country's development. OECD countries were divided into clusters based on the indicators: the share of Bachelor program graduates in total number of graduates, the share of Master program graduates in total number of graduates, the share of short-term courses graduates in the total number of graduates, the share of PhD graduates in the total number of graduates. The author developed multi-factor regression models of the dependence of the level of employment on the weight and composition of the population with tertiary education diplomas. The models were compared among the clusters, and specific characteristics of the labor market patterns were established. Comparative assessments of the efficiency of higher education were accomplished.

C0442: **Uncovering well-being patterns: An archetypal analysis of development and happiness**
Presenter: **Claudeline Cellan**, Bangko Sentral ng Pilipinas, Philippines

Development is multifaceted, and recent efforts to accelerate growth across all dimensions are gaining traction. The Human Development Index (HDI) evaluates development beyond income by incorporating life expectancy, education performance and income per capita. Meanwhile, the World Happiness Report produces happiness scores based on self-assessed well-being. The aim is to explore "patterns of well-being" by merging HDI metrics with happiness scores. Archetypal analysis identifies archetypes by examining the extremities of the multivariate data to uncover realistically observable patterns. The analysis revealed three main archetypes: (1) a country with high ratings in both human development indicators and happiness, (2) one with low ratings in both areas and (3) a country with high ratings in development indicators but low income, yet high happiness ratings. While the first two are expected, the third pattern is particularly intriguing, as it encompasses the majority of countries. Tracking the degree of belonging of the countries provides insights into the changing landscape of a country's development and happiness, particularly post-COVID-19 pandemic.

C0453: **Detecting and understanding wash trading on cryptocurrency exchanges**
Presenter: **Jan Sila**, UTIA AV CR, v.v.i., Czech Republic
Co-authors: Ladislav Kristoufek, Jiri Kukacka, Evzen Kocenda

Wash trading on cryptocurrency exchanges is investigated, and the behavioral and socioeconomic factors are considered to influencing wash trading volumes. The findings show that wash trading is common on unregulated exchanges across asset classes, with exchanges inflating volumes by 50 per cent at times. These artificial volumes skew prices and boost exchange rankings. Strong correlations are found between wash trading volumes and external indicators such as Google search trends and return momentum, implying that market sentiment and socioeconomic conditions motivate exchanges to increase wash trading volumes as they compete for retail crypto traders by improving their rankings.

Wednesday 28.08.2024 11:00 - 12:30 Parallel Session F – COMPSTAT2024

CI010 Room 45 BAYESIAN COMPUTATIONAL METHODS Chair: Mattias Villani

C0332: **Generative models and approximate Bayesian inference**
Presenter: **Christian Andersson Naesseth**, University of Amsterdam, Netherlands

Generative models have taken the world by storm. Generative modelling, or generative AI, is the task of constructing an approximation to the data-generating process in the form of a probability distribution. In the context of text, for example, in large language models, the distribution is over words (or tokens), whereas for images, it is an approximate probability distribution over pixel values. The similarities, connections, and potential synergies between generative AI and approximate Bayesian inference are discussed.

C0350: **Insufficient Gibbs sampling**
Presenter: **Christian Robert**, Universite Paris-Dauphine, France
Co-authors: Robin Ryder, Antoine Luciano

In some applied scenarios, the availability of complete data is restricted, often due to privacy concerns; only aggregated, robust and inefficient statistics derived from the data are made accessible. These robust statistics are not sufficient, but they demonstrate reduced sensitivity to outliers and offer enhanced data protection due to their higher breakdown point. A parametric framework is considered, and a method to sample from the posterior distribution of parameters conditioned on various robust and inefficient statistics is proposed: specifically, the pairs (median, MAD) or (median, IQR), or a collection of quantiles. The approach leverages a Gibbs sampler and simulates latent augmented data, which facilitates simulation from the posterior distribution of parameters belonging to specific families of distributions. A by-product of these samples from the joint posterior distribution of parameters and data given the observed statistics is that it can estimate Bayes factors based on observed statistics via bridge sampling. The limitations of the proposed methods are validated and outlined through toy examples and an application to real-world income data.

C0504: **Bayesian inference of pharmaceutical models with Pumas: A showcase of Julia for high-performance interactive computing**
Presenter: **David Widmann**, Pumas-AI, USA, Sweden

Pharmacometric models are mathematical models that describe relationships between patient characteristics, administered doses, drug concentrations, and observed biomarkers. Understanding such relationships is important for drug development and (individualized) drug therapies. Pumas is a modern platform for pharmaceutical modeling and simulation that builds on a domain specific modeling framework in the Julia programming language to achieve high-performance analytics. The aim is to present its use for Bayesian inference of pharmaceutical models and its integration with and connection to the Julia ecosystem.

CO109 Room 051 COMPUTATIONAL AND STATISTICAL METHODS IN CLINICAL RESEARCH Chair: Maria del Carmen Pardo

C0366: **Issues with the R-squared for the evaluation of polygenic prediction models across diverse ancestries**
Presenter: **Christian Staerk**, IUF - Leibniz Research Institute for Environmental Medicine, Germany
Co-authors: Hannah Klinkhammer, Tobias Wistuba, Carlo Maj, Andreas Mayr

Polygenic risk scores (PRS) quantify genetic predispositions for traits and clinical outcomes based on genotype data. For effective personalized risk assessment, polygenic prediction models should generalize well across diverse ancestries. However, the commonly used R-squared measure is not unambiguously defined for test data, complicating the assessment of prediction accuracy of PRS models and the interpretation of results. Recent scalable regression methods are applied, including statistical boosting on large-scale individual-level genotype data from the UK Biobank, and three R-squared definitions are compared for evaluating the predictive performance of PRS models on different populations. It is found that the choice of R-squared definition considerably affects the results: while the squared correlation between predicted and observed phenotypes always stays between 0 and 1, R-squared definitions incorporating the squared prediction error can yield negative values, particularly for miscalibrated prediction models. It is argued that the choice of the most appropriate definition of the R-squared depends on the aim of the PRS analysis, i.e., whether the PRS should be mainly used for risk stratification in a given cohort or also for the prediction of continuous traits for individual risk assessment. Further research is needed to develop and evaluate well-calibrated polygenic models across diverse ancestries in clinical practice.

C0288: **Estimating the cure rate in a mixture cure model using presmoothing**
Presenter: **Maria Amalia Jacome Pumar**, Universidade da Coruna, Spain
Co-authors: Ana Lopez-Cheda, Samuel Saavedra

The mixture cure model in survival analysis has received large and growing attention in the last few decades. For diagnostic and prognostic purposes, an important aspect of the mixture cure model is the estimation of the probability that an individual is cured, that is, belongs to the cured component of the population. There is no reason to believe that the cure rate is always monotone, let alone that it is logistic. Hence, we consider the nonparametric estimator for the cure rate function. It is given by the Kaplan-Meier survival estimator (or the generalized Kaplan-Meier estimator, with covariates) evaluated at the largest uncensored time. Presmoothing has been shown to improve survival function estimation, by replacing the indicators of no censoring with some preliminary nonparametric estimator of the conditional probability of uncensoring. The effect of presmoothing in the estimation of the cure rate will be studied, and the resulting methods will be applied to a real medical database.

C0326: **Smoothing spline density estimation from doubly truncated data**
Presenter: **Jacobo de Una-Alvarez**, University of Vigo, Spain

Smoothing splines can be introduced as the solution to a penalized maximum likelihood problem in a reproducing kernel Hilbert space. In the setting of density estimation, they provide a flexible, nonparametric approach to approximate the target, achieving a compromise between bias and variance in an automatic way. Doubly truncated data are often encountered in survival analysis, epidemiology, reliability engineering, economics or astronomy, among other fields. They appear in particular with interval sampling, when the sampled units are those for which the event of interest occurs between two particular dates, a sampling mechanism that is ubiquitous, for instance, in clinical research. Estimation from doubly truncated data requires proper corrections for the sampling bias. Smoothing splines are introduced and investigated for density estimation in the presence of double truncation. Through Monte Carlo experiments, the relative benefits of smoothing splines compared to other popular nonparametric approaches, such as kernel density estimation, are illustrated. Real data illustrations are provided.

C0224: **Relationships between summary ROC indices and overlap coefficients**
Presenter: **Maria del Carmen Pardo**, Complutense University of Madrid, Spain
Co-authors: Alba Franco-Pereira

One summary ROC index that was very recently studied is the length of the ROC curve. The length successfully captures useful biomarkers, which would have been rejected using standard approaches. Lastly, overlap measures such as Weitzman's measure or Bhattacharyya coefficient have been studied as measures of medical diagnostic test accuracy. We put all these measures under one unifying framework, and some theoretical results regarding their interrelationships have been proved. Both parametric and nonparametric estimators are proposed for these measures. A simulation study is conducted to evaluate the performance of our approaches. Finally, an illustrative example is presented from a study on cancer biomarkers.

CO096 Room 052 OPTIMAL EXPERIMENTAL DESIGN AND APPLICATIONS Chair: Frank Miller

C0266: **Discriminating among several random effects models**
Presenter: **Chiara Tommasi**, University of Milan, Italy

Random effects models are largely applied across all disciplines, particularly in clinical studies and biosciences. The focus is on the design issue of optimally discriminating among several random effects models using the Kullback-Leibler divergence (KL) criterion. A theoretical result proves that the KL criterion leads to classical T optimality based on a different inner product. A closed-form expression for the minimum Kullback-Leibler divergence is provided, which makes much easier the implementation of the algorithms to find out optimal designs. Finally, two examples show how to apply the proposed methodology. The first application concerns discrimination among fractional polynomials with a single continuous variable; the latter identifies the best design to discriminate among several multi-factorial random effects models.

C0329: **The polytope of optimal approximate designs: Extending the selection of informative experiments**
Presenter: **Radoslav Harman**, Comenius University Bratislava, Slovakia
Co-authors: Lenka Filova, Samuel Rosa

Consider the problem of constructing an experimental design that is optimal for a given statistical model with respect to a chosen criterion. To address this problem, the literature usually provides a single solution. Sometimes, however, there exists a rich set of optimal designs, and the knowledge of this set can lead to substantially greater freedom in selecting an appropriate experiment. It is demonstrated that the set of all optimal approximate designs generally corresponds to a polytope. Particularly important elements of the polytope are its vertices, which are called vertex optimal designs. It is proven that the vertex optimal designs possess unique properties, such as small supports, and outline strategies for how they can facilitate the construction of suitable experiments. Moreover, it is shown that for a variety of situations, it is possible to construct the vertex optimal designs with the assistance of a computer by employing error-free rational-arithmetic calculations. With this approach, the polytope of optimal designs is determined for several common multifactor regression models, thereby extending the theoretical knowledge and the choice of informative experiments for these models.

C0192: **Optimal adaptive two-stage designs**
Presenter: **Maximilian Pilz**, University of Heidelberg, Germany
Co-authors: Meinhard Kieser

To conduct a clinical trial, the adequate choice of the required sample size is a crucial decision. When too many patients are recruited, they are exposed to an unnecessary risk of a useless or even harmful intervention. Including too few patients, however, raises the risk that potential underlying effects may not be detected with sufficient probability. Adaptive two-stage designs offer an attractive option to improve the sample size determination procedure. An interim analysis is performed during the ongoing trial, during which the sample size may be adjusted according to the data observed so far. To plan an adaptive design, one has to choose the sample size based on the interim analysis and the sample size adjustment rule. We present how the determination of an adaptive clinical trial design can be formulated as an optimization problem and how this problem is solved. Different properties of an adaptive design are discussed, and how they can be included jointly in the optimization problem is demonstrated. We also discuss the optimization of specific designs, e.g., group-sequential designs or designs based on the inverse normal combination method. We conclude with a multicriteria optimization view on the choice of an adaptive two-stage design.

C0371: **Optimizing test item calibration - with application to the Swedish national mathematics test**
Presenter: **Ellinor Fackle-Fornius**, Department of Statistics, Sweden
Co-authors: Frank Miller

For large-scale achievement tests, like national tests in school, test items need to be calibrated before being implemented in the test. Item calibration is the process of estimating item parameters such as difficulty and discrimination through pretesting. It's crucial to estimate the item parameters with as high precision as possible, as it influences the test quality and the accuracy of ability estimates for the examinees. Instead of randomly allocating calibration items to examinees, optimal design theory is utilized to allocate items to examinees in an optimal way based on their individual ability levels. An ability-matched item allocation method is employed, tailored to handle item calibration conducted in large groups in parallel with items of mixed formats, a common scenario for many large achievement tests. The optimal design, IRT analyses, and results of a real calibration study conducted for the national test in mathematics in Sweden are presented.

CO080 Room 43 ADVANCES IN DISTRIBUTIONAL REGRESSION Chair: Thomas Kneib

C0307: **Distributional regression for lung function of cystic fibrosis patients with a special focus on center-specific effects**
Presenter: **Elisabeth Bergherr**, Georg-August-Univeritat Gottingen, Germany
Co-authors: Colin Griesbach, Marisa Lane

Rapid lung function decline is a severe problem for cystic fibrosis patients throughout their whole life. We have access to the data from the German cystic fibrosis registry, which includes thousands of patients with hundreds of thousands of observations repeatedly over each year and hundreds of variables, like sociodemographic information, biomarkers, but also gene expression data. We plan to estimate a prediction model for the lung volume measured by the %FEV1-value. The latter is one of the key indicators for healthy functionality of the lung. Since not only the expectation of the volume but also the variation is of major importance for the patients, a Gaussian distributional regression model will be used. The vast amount of possible explanatory variables calls for a strong selection algorithm, which is one of the key features of gradient boosting. We will develop a way of including the information on the center in which the individual patients are treated. The algorithm will hence either detect a spatial pattern, or account for the center specific variation in terms of (possibly clustered) random effects.

C0294: **Enhanced variable selection for boosting sparser and less complex models in distributional regression**
Presenter: **Andreas Mayr**, University of Bonn, Germany
Co-authors: Annika Stroemer, Nadja Klein, Christian Staerk, Guillermo Briseno Sanchez

Variable selection is already a challenge in classical regression, but in the context of distributional regression, where we model different parameters of the conditional distribution, it becomes even more pressing. An automated approach to deal with this is statistical boosting. These kinds of algorithms are able to select the most informative variables while fitting the corresponding models. Unfortunately, in many practical applications with a low to medium number of predictor variables, they have a tendency towards false positives. This does not necessarily harm prediction accuracy, as the falsely selected variables often have only a negligible impact on the final model. However, this behavior hinders the interpretation of the models. As the general aim of statistical modelling is to utilize models as complex as necessary but also as simple as possible, we investigate different approaches to enhance the variable selection properties of boosting, focusing on probing, stability selection, and a recent deselection approach. We will illustrate the effect of these approaches on variable selection and prediction accuracy with different model classes, including copula regression for multivariate distributional regression.

C0320: **Neural distributional regression models**
Presenter: **Benjamin Saefken**, Clausthal University of Technology, Germany

Users of neural regression models have a tendency to neglect distributional aspects of the data. This is often because appropriate frameworks are not available. Neural Additive Models for Location, Scale and Shape (NAMLSS) provide a framework that allows the modelling of distributional aspects of the target data. This approach offers many advantages for the user, such as accurate prediction intervals. In comparison to classic statistical regression models, their neural counterparts allow for the possibility of incorporating non-tabular data. However, for a statistician, this is only of interest if it is done in an interpretable fashion. Otherwise, inferential techniques are not sensible. An approach for specifically incorporating images in a comprehensible way is proposed based on embedding spaces. Results on particular applications are promising. The downside is that the method is not easily generalizable to other settings and data sets and not as mathematically rigid as common regression models for tabular data.

C0465: **Graphical conditional transformation models**
Presenter: **Matthias Herp**, Georg-August-Universitaet Goettingen, Germany
Co-authors: Johannes Brachem, Thomas Kneib, Michael Altenbuchinger

Graphical conditional transformation models (GCTMs) are proposed as a novel approach to effectively model multivariate regression data with intricate marginals and complex dependency structures non-parametrically while maintaining interpretability through the identification of conditional independencies. Multivariate conditional transformation models (MCTMs) are built upon a combination of marginal conditional transformation models(CTMs) with a conditional Gaussian copula by exchanging the copula with a custom-designed transformation. This has two major advantages. First, the GCTM can learn more complex interdependencies by using penalised splines, which also provide an efficient regularisation scheme. Second, it shows how to regularise the GCTM with a lasso penalty towards pairwise conditional independencies similar to Gaussian graphical models (GGMs). The robustness and effectiveness of the model are validated through simulations, demonstrating its ability to accurately learn parametric vine copulas, identify conditional independencies, and incorporate covariates. In addition, the model is applied to two real-world data sets. For a benchmark astrophysics data set, the model is demonstrated to compare favorably with non-parametric Vine Copulas in learning complex multivariate distributions. Similarly, in a genomics data set, it is shown that the model learns sparse undirected graphs, outperforming GGMs with transformed marginals.

CO108 **Room 44** **MODELING COMPLEX DATA WITH DEPENDENCIES** **Chair: Sara Taskinen**

C0173: **Adjusted predictions in generalized estimation equations**
Presenter: **Francis Hui**, The Australian National University, Australia
Co-authors: Muller Samuel, Alan Welsh

Generalized estimating equations (GEEs) is a popular regression approach that requires specification of the first two marginal moments of the data, along with a working correlation matrix capturing the covariation between responses e.g., temporal correlations within clusters in longitudinal data. The majority of research and application of GEEs has focused on the estimation and inference of regression coefficients in the marginal mean. When it comes to prediction using GEEs, practitioners often simply and quite understandably also base it on the regression model characterizing the marginal mean. We propose a simple adjustment to predictions in GEEs based on utilizing information in the assumed working correlation matrix. By viewing the GEE from the perspective of solving a working linear model, we borrow ideas from universal kriging to construct a predictor that leverages temporal correlations between the new and current observations within the same cluster. We establish some theoretical conditions for the proposed adjusted GEE predictor to outperform the standard unadjusted predictor. Simulations show even when we misspecify the working correlation, adjusted GEE predictors (combined with an information criterion for choosing the working correlation matrix) can improve the predictive performance of standard GEE predictors as well as the so-called oracle GEE predictor using all observations.

C0325: **Fast fitting of phylogenetic random effect models**
Presenter: **Bert van der Veen**, Norwegian University of Science and Technology, Norway
Co-authors: Robert OHara

Ecologists survey locations in space or time to collect data on the presence or abundance of species. The models fitted to such data are used to assess the impact of changes in environmental conditions on species and potentially make recommendations for conservation purposes. Often, there are few non-zero observations for many of the species, so adding an extra source of information would help to successfully estimate the parameters in the models. One way to do that is to add information on the relatedness of species via a Phylogenetic tree. This allows us to borrow information from species that occur more frequently in order to determine the response of less abundant species that share a similar evolutionary ancestry. We present an implementation of Phylogenetic random effects models in the gllvm R package, which uses Variational Approximations (VA) for estimation. The key to a fast approximation is to reduce the number of VA parameters as much as possible, which we do by applying a matrix normal structure for the VA distribution and combining it with a reduced rank approximation. We also apply a Nearest Neighbor approximation to the inverse for the Phylogenetic covariance matrix, and use Template Model Builder for parallel computations.

C0245: **Cubble: An R Package for organizing and wrangling multivariate spatio-temporal data**
Presenter: **Huize Zhang**, University of Texas at Austin, United States
Co-authors: Di Cook, Ursula Laa, Nicolas Langrene, Patricia Menendez

Multivariate spatio-temporal data have a spatial component referring to the location of each observation, a temporal component recorded at regular or irregular time intervals, and multiple variables measured at each spatial and temporal value. Often, such data are fragmented, reflecting a common practice of focusing on either spatial or temporal aspects separately. This fragmentation makes it difficult to handle them coherently and comprehensively. A new data structure is introduced to facilitate the study of different portions or combinations of spatio-temporal data for exploratory data analysis. The proposed structure, implemented in the R package, cubble, organizes spatial and temporal variables as two facets of a single data object, allowing them to be wrangled separately or combined while ensuring synchronization. Examples of creating glyph maps will be provided to visualize weather station data with cubble.

C0372: **Advances in complex-valued covariance modeling**
Presenter: **Sandra De Iaco**, University of Salento, Italy

As in the real case, the development of the complex formalism in a spatial or spatiotemporal context and the construction of some new classes of complex covariance models are of sure interest to the scientific community partly due to the ongoing explosion in the availability of vector observations. Moreover, taking into account that monitoring environmental networks provides datasets for multiple scalar and vector variables, some advances in complex covariance modelling are also required, especially in some applied fields, such as in the area of the marine environment. Indeed, this kind of study deserves attention for its connections with the transport of nutrients and pollutants, which might influence marine plants and animals and biodiversity. Then, an application to oceanographic data is presented.

CC059 **Room 001** **MULTIVARIATE DATA ANALYSIS** **Chair: Bojana Milosevic**

C0160: **Simple procedure to estimate a structural equation model with latent variables**
Presenter: **Zouhair El Hadri**, Mohammed V University, Morocco

Structural equation modelling is a sophisticated approach used to analyse complex models with both observed variables, called manifest variables or indicators, and unobserved variables, called latent variables, factors, components or constructs. The aim is to introduce a simple procedure to estimate the parameters associated with structural equation models. The proposed procedure is the extension of the procedure introduced recently for path analysis models (models without latent variables). First, we show that the covariance matrix implied by the model is affine with respect to each parameter. Second, we use this affinity property to build an iterative procedure to estimate the parameters. Third, we provide proof of the monotony convergence of the function minimized in the said procedure. Finally, we provide illustrative real data and numerical simulations.

C0424: **Supervised dimension reduction for instrumental variables estimation with some invalid instruments**
Presenter: **Kei Tsubotani**, Graduate School of Doshisha University, Japan
Co-authors: Jun Tsuchida, Hiroshi Yadohisa

The instrumental variables (IVs) method is used to estimate treatment effects without bias when an unobserved confounder exists. This method relies on valid IVs that satisfy three assumptions: relevance, exclusion restriction, and independence. However, identifying valid IVs can be challenging without sufficient knowledge of the domain. To select valid IVs when the candidate IVs include invalid IVs that do not meet the exclusion restriction (i.e., variables that directly affect the outcome), several methods have been proposed by applying variable selection techniques commonly used in regression analysis. Meanwhile, supervised dimension reduction can be used to extract valid IVs because the subspace estimation method includes variable selection techniques. A method for estimating treatment effects using supervised dimension reduction is proposed under an assumed situation. Numerical experiments are conducted to evaluate the performance of the IV method using supervised dimension reduction methods from the viewpoint of the number of valid IVs and the strength of the relationship between the outcome and invalid IVs. The results reveal that the performance of the supervised dimension reduction methods was superior to that of the variable selection methods when the number of valid IVs was small.

C0443: **Tests for the multivariate skew-normal distribution based on data transformations**
Presenter: **Aurora Monter-Pozos**, Colegio de Postgraduados, Mexico
Co-authors: Elizabeth Gonzalez-Estrada

The parametric statistical inference relies on the assumption that the data are a random sample from a population that has a given probability distribution. This kind of inference is valid only if the probability distribution used really explains the probabilistic behaviour of the data. A probability distribution that has gained importance in the last decades is the multivariate skew normal (MSN) distribution, which extends the normal one by incorporating a shape parameter. It provides probability models for datasets showing moderate degrees of skewness. Goodness-of-fit tests for the MSN distribution are proposed based on the canonical transformation and an additional transformation to multivariate normality. Then, the problem of testing the null hypothesis that a random sample follows an MSN distribution with unknown parameters is reduced to testing that the sample comes from a multivariate Gaussian distribution. Simulation results show that the proposed tests have desirable properties and are competitive against existing tests for the same problem. A real data set is analyzed in order to illustrate the usefulness of the tests.

C0154: **A sequential method to search for multiple outliers in multivariate data**
Presenter: **Trijya Singh**, Le Moyne College, Syracuse, NY, United States

In usual multivariate analysis methods such as principal components, discriminant analysis and so on, the sample mean vector and covariance matrix are utilized. These can be strongly affected by the presence of only a few outliers. The problem of detecting outliers in multivariate data sets can be difficult because classical methods based on Mahalanobis distances may work well for identifying scattered outliers but perform poorly in the case of multiple clustered outliers. Methods based on robust Mahalanobis distances also do not perform well when the fraction of contamination is high and can also be computationally expensive. A method of detecting multiple outliers in multivariate data is proposed, which involves sequential testing of outliers and utilizes the leave-one-out approach at many stages. The proposed method is applied to a well-known data set, and it is shown that it is marginally better to first obtain a clean sample to estimate the mean vector and covariance matrix and then apply classically efficient methods rather than using inefficient robust rules for estimation and subsequent outlier detection.

CC137 Room 050 PRACTICAL INSIGHTS IN SPATIAL STATISTICS Chair: Claudia Kirch

C0421: **Reliability evaluation of regions within hotspot clusters using hierarchical structure of spatial data**
Presenter: **Yusuke Takemura**, Kyoto Womenś University, Japan
Co-authors: Fumio Ishioka, Koji Kurihara

In data such as the observed number of deaths due to infectious diseases in each region, there can be areas where the mortality risk is significantly higher than in the surrounding areas. These areas are referred to as hotspot clusters. Several methods utilizing spatial scan statistics have been proposed to detect clusters, and these methods have been widely used in fields such as epidemiology. However, in cluster detection using spatial scan statistics, there are regions that are erroneously detected as clusters due to slight fluctuations in observed values. Therefore, it is important to evaluate whether the regions detected as clusters should truly be included in clusters. To address this issue, the focus is on echelon analysis, a method for representing the hierarchical structure of spatial data. This method topologically classifies each region based on univariate values such as mortality rates, positioning regions with higher values in the higher hierarchy. A method is introduced for evaluating reliability by clarifying where regions included in detected clusters are located within the hierarchical structure obtained through echelon analysis.

C0422: **A novel method for spatial cluster detection in continuous data**
Presenter: **Fumio Ishioka**, Okayama University, Japan
Co-authors: Yusuke Takemura, Koji Kurihara

Spatial phenomena concentrated in specific regions, such as the mortality rate of certain diseases across municipalities, are referred to as "clusters". In recent years, in fields like spatial epidemiology, spatial scan statistics have been widely employed to explore specific regions and evaluate the presence of clusters using likelihood-based methods. This test comprises two elements: 1) a statistical model and 2) a scanning method, which are combined for analysis. However, current spatial scan statistics predominantly employ Poisson models for counting data (discrete values), such as the number of disease cases or traffic accidents, on the statistical model front. Alternatively, while weighted Normal models accommodating regional variations are suggested for continuous data, circular scan methods remain predominant. These methods mainly focus on concentric scanning of regions, presenting difficulties in accurately identifying non-circular clusters, like those following rivers or roads. Therefore, to detect clusters of arbitrary shapes when analyzing continuous values, it is endeavored to apply the echelon scan method developed to the scanning approach. Furthermore, the aim is to validate how the application of this method influences the accuracy of cluster detection.

C0462: **Geographically weighted principal components analysis and variography for environmental variables**
Presenter: **Monica Palma**, University of Salento, Italy
Co-authors: Sabrina Maggio, Giuseppina Giungato

Several statistical techniques of multivariate analysis, such as principal component analysis (PCA), allow the researchers to build very simple way composite indicators measuring the phenomenon under study. In social and environmental sciences, it is very common to synthesize the variables of interest in a unique indicator suitable to describe the variables at hand and which could be used by policymakers as decision support. In the presence of a multivariate geo-referenced data set, the classical multivariate techniques are not at all adequate to define composite indicators. In this case, the geographically weighted PCA (GWPCA) is a valid approach to constructing spatial composite indicators, taking into account the multivariate spatial dependence of the study variables. Recently, a new approach to choosing one of the GWPCAs parameters, i.e., the bandwidth of the kernel weighting function, has been proposed in a recent study and applied in a socio-economic context. The novel approach for GWPCA is used to construct a spatial composite indicator of urban air quality over a risk area, and the appropriateness of the found indicator in estimating the environmental quality at un-sampled locations is also discussed.

C0468: **Twofold nested error regression models with data-driven transformation**

Presenter: **Rachael Katwa Kyalo**, Otto-Friedrich-Universitaet Bamberg, Germany
Co-authors: Timo Schmid, Nora Wuerz

Small area estimation effectively addresses the issue of small sample sizes within subpopulations. Typically, the target population is divided into multiple nested hierarchical levels, such as counties and sub-counties. A twofold nested error regression model with random effects captures the variability across these levels. For estimating non-linear indicators like poverty measures, the twofold EBP model can be used, which relies on normality assumptions of the error terms - a condition often unmet in real data applications. The twofold nested error regression model is enhanced by incorporating a data-driven transformation, improving the model's robustness and flexibility. MSE estimation is performed using resampling methods. Model-based simulations compare the proposed model's performance with onefold EBP methods that include either area or sub-area effects. Results show that the proposed twofold EBP method adapts to the distribution shape, providing more efficient estimates than a fixed logarithmic transformation or no transformation. Finally, the twofold EBP with data-driven transformation is used to generate poverty estimates for rural and urban regions within Kenyan counties, offering a more nuanced and accurate assessment of poverty levels.

Wednesday 28.08.2024 14:00 - 16:00 Parallel Session G – COMPSTAT2024

CV060 Room 050 MULTIVARIATE DATA ANALYSIS AND GRAPHICAL MODELS Chair: George Karabatsos

C0383: **A small-sigma approximation for LIML and FLIML to estimation bias in the dynamic simultaneous equation model**
Presenter: **Emma Iglesias**, University of A Coruna (SPAIN), Spain
Co-authors: Garry Phillips

Small-sigma approximations for estimator bias in the dynamic simultaneous equation model (DSEM) have previously been presented for OLS and 2SLS in the literature, where both dynamic and simultaneity bias components are present. FLILM has been shown to be useful in removing the bias of order T 1 in the static SEM (SSEM), but its performance is unknown in the DSEM. A bias approximation is provided for FLIML that shows that only a dynamic bias component exists and the simultaneity component disappears. A bias approximation is also included for LIML. It is shown that the FLIML estimator is the only one that removes the simultaneity bias component of common k-class estimators in the DSEM, which, surprisingly, all have the same dynamic component. Theoretical and simulation evidence also shows that FLIML works best overall.

C0430: **A Bayesian approach to ensemble clustering**
Presenter: **Elena Ballante**, Department of Political and Social Sciences, University of Pavia, Italy
Co-authors: Federico Maria Quetti, Silvia Figini

In the context of ensemble clustering, little attention has been given to the integration of conventional bootstrap methodologies within clustering frameworks. The aim is to bridge this gap by introducing an innovative approach that enhances clustering techniques through the application of Bayesian bootstrap techniques. The method leverages insights gleaned from bootstrap resampling, incorporating a Gaussian mixture as the prior distribution of group densities. The methodology comprises two steps. Initially, the Efron bootstrap method is employed to robustly estimate the parameters of the prior distribution from the available data. Then, a proper Bayesian bootstrap is applied to resample from a mixture of the prior distribution and the empirical distribution of the data. The exploitation of prior knowledge jointly with observed data fosters a synergistic approach, enhancing the adaptability and robustness of the clustering process. Moreover, the application of bootstrap naturally leads to a fuzzy clustering interpretation of the results, providing better interpretability of the algorithm as well as ideas for future advancements. The proposed methodology also shows promising results for the determination of the optimal number of clusters. Varying the number of clusters and the variance of the prior distribution, the method offers fundamental insights into the underlying cluster structure of the data, even in scenarios characterized by high dimensionality of the data.

C0408: **Response prediction with convergence guarantees in multiple random graphs on unknown manifolds**
Presenter: **Aranyak Acharyya**, Johns Hopkins University, United States
Co-authors: Jesus Arroyo, Michael Clayton, Marta Zlatic, Youngser Park, Carey Priebe

The popularity of random graphs has increased in recent times owing to their applicability in analyzing network data arising from various spheres of real life, including neuroscience, biology and social studies. The model involves a collection of graphs with a shared structure on a common set of nodes, where some of the graphs are associated with responses. Assuming that each graph corresponds to a point on a one-dimensional manifold in higher dimensional ambient space, a technique that predicts the response is proposed at an unlabeled graph by exploiting the underlying manifold structure, which is unknown and hence estimated from the data. Convergence guarantees for the method are established, and its performance is demonstrated on simulated data.

C0486: **Structure learning for zero-inflated counts, with an application to single-cell RNA sequencing data**
Presenter: **Thi Kim Hue Nguyen**, University of Padova, Italy
Co-authors: Monica Chiogna, Davide Risso

In recent years, a growing interest has developed around the problem of retrieving, starting from observed data, the structure of graphs representing relationships among variables of interest. In fact, the reconstruction of a graphical model, known as structure learning, traces back to the beginning of the nineties, and a vast amount of literature exists that considers the problem from various perspectives within both frequentist and Bayesian approaches. However, molecular biology applications have played a central role in renewing interest in structure learning. In this field, the abundance of data with increasingly large sample sizes, driven by novel high-throughput technologies, has opened the door for the development and application of structure learning methods, in particular, applied to the estimation of gene regulatory or gene association networks. These, however, are challenging applications since the data consists of high-dimensional counts with high variance and over-abundance of zeros. A general framework is presented for learning the structure of a graph from single-cell RNA-seq data based on the zero-inflated negative binomial distribution. The approach is demonstrated with simulations to retrieve the structure of a graph in various settings, and the utility of the approach is shown on real data.

C0512: **A novel computational methodology for clinical characteristic predictive gene network estimation**
Presenter: **Heewon Park**, Sungshin Womenś University, Japan

We propose a novel computational methodology for clinical characteristic (e.g., drug sensitivity of cell lines) predictive gene network estimation, called a PredictiveNetwork. The objective function of the PredictiveNetwork consists of loss functions for gene network estimation and prediction, and thus we can estimate gene network and predict clinical characteristic, simultaneously. It implies that the network is estimated to be optimized for not only network estimation but also explain the clinical characteristic, thus we can identify clinical characteristic prediction specific molecular interplays. We extend the PredictiveNetwork to network-based classification and develop a Gene regulatory network-based classifier (GRN-classifier) that estimates the gene network to minimize errors for both network estimation and classification of cell lines, in line with the PredictiveNetwork. The proposed strategies are applied to gastric cancer drugs response predictive network estimation and related marker identification, especially we focus on drug resistance molecular interplays identification. The PredictiveNetwork is applied to gastric cancer drugs response predictive network estimation, and GRN-classifier is applied to classify 5-FU -sensitive/resistant and 5-FU target/non-target cell-lines. Our analysis results suggest that active regulatory system between AKR family is a crucial clue to uncover mechanism of acquired gastric cancer drug resistance.

CO114 Room 43 STATISTICS ON SHAPES AND MANIFOLDS Chair: Joern Schulz

C0258: **Averaging symmetric positive-definite matrices on the space of eigen-decompositions**
Presenter: **Sungkyu Jung**, Seoul National University, Korea, South

Extensions of Fréchet means for random objects in the space of symmetric positive-definite matrices are studied using the scaling-rotation geometric framework. The scaling-rotation framework is designed to enjoy a clearer interpretation of the changes in random ellipsoids in terms of scaling and rotation. The framework has been beneficial in smoothing coarse diffusion tensor imaging. We formally define the scaling-rotation (SR) mean set as the set of Fréchet means with respect to the scaling-rotation distance. Since computing such means requires a discrete optimization, we instead define the partial scaling-rotation (PSR) mean set lying on the space of eigen-decompositions as a proxy for the SR mean set, which is easier to compute and often coincides with the SR mean set. Even though the PSR mean is never unique, we reveal sufficient conditions for the mean to be unique up to the action of a certain group. On a theoretical side, a procedure is illustrated for deriving strong consistency and a central limit theorem for M-estimators, defined in a non-metric and stratified space. In an application to multivariate tensor-based morphometry, we demonstrate that a two-group test using the proposed PSR means has greater power than using the usual Log-Euclidean geometric framework for symmetric positive-definite matrices.

C0308: **Bi-invariant dissimilarity measures for sample distributions in lie groups**
Presenter: **Christoph von Tycowicz**, Zuse Institute Berlin, Germany
Co-authors: Hans-Christian Hege, Martin Hanik

Data sets sampled in Lie groups are widespread, and as with multivariate data, it is important for many applications to assess the differences between the sets in terms of their distributions. Indices for this task are usually derived by considering the Lie group as a Riemannian manifold. Then, however, compatibility with the group operation is guaranteed only if a bi-invariant metric exists, which is not the case for most non-compact and non-commutative groups. By using an affine connection structure instead, we will obtain bi-invariant generalizations of well-known dissimilarity measures: the Hotelling T^2 statistic, the Bhattacharyya distance, and the Hellinger distance. Each of the dissimilarity measures matches its multivariate counterpart for Euclidean data and is translation-invariant, so that biases, e.g., through an arbitrary choice of reference, are avoided. We will examine the potential of these dissimilarity measures by performing group tests on knee configurations and epidemiological shape data.

C0226: **Barycentric subspace analysis of a set of graphs**
Presenter: **Elodie Maignant**, Zuse Institute Berlin, Germany
Co-authors: Xavier Pennec, Anna Calissano

In the context of statistical analysis of populations of graphs, unlabeled graphs are usually modeled as equivalence classes under the action of permutations on nodes or, equivalently, under the action by conjugation of permutation matrices on adjacency matrices. Such a model relies, however, on the combinatorial problem of graph matching and is therefore computationally limited. As a relaxation of the previous action, we introduce a new framework where graphs are modeled in the quotient space resulting from the action of rotations on the set of adjacency matrices. Now, beyond the idea of relaxation, our approach takes on a natural interpretation from the point of view of spectral graph theory, that is, the study of graphs through their spectrum, a descriptor that has been proven to encode numerous structural properties. Indeed, the action of rotations by conjugation preserves the eigenvalues of a graph (represented by its adjacency matrix), and each resulting equivalence class is exactly the set of all the graphs with a given spectrum, also called cospectral graphs. We unveil the very particular and surprisingly simple geometry of these quotient spaces. Then, within such model, we investigate Barycentric Subspace Analysis (BSA), a non-Euclidean dimensionality reduction method. Through several examples, we illustrate that BSA is a powerful dimensionality reduction tool with great interpretability, particularly when it is used to analyze a set of graphs.

C0363: **Maximum entropy ensemble refinement**
Presenter: **Benjamin Eltzner**, Max Planck Institute for Multidisciplinary Sciences, Germany
Co-authors: Bert de Groot, Michael Habeck, Daniel Rudolf, Julian Hofstadler

In some cases, features measured from a protein ensemble, like atom distances, are not recovered on average in molecular dynamics simulations. The problem is approached from the maximum entropy point of view. The problem then presents as a Bayesian inference problem with unknown likelihood normalization, a so-called doubly intractable problem. This type of problem requires sophisticated two-step Monte Carlo methods. The focus is on NOE measurements, where the measured peak intensities are proportional to the inverse sixth power of atomic distance. Ensembles derived using the corresponding maximum entropy energy terms in a modified molecular dynamics simulation show a wider variety of structures than ensembles derived by other refinement methods while still satisfying measured feature bounds on average.

C0323: **Statistics on locally parametrized shapes via discrete swept skeletal representations**
Presenter: **Joern Schulz**, University of Stavanger, Norway
Co-authors: Mohsen Taheri Shalmani

Effective statistical shape analysis, such as detecting local dissimilarities of image objects, depends on the underlying shape representation and how the representation establishes local correspondence between the objects in a population. In opposite to most shape representations that are based either on non-invariant spatial geometrical object properties (GOPs) or on extrinsic GOPs, we propose a novel skeletal shape representation by defining local coordinate systems (fitted frames) at each location on the object skeletal and thereby defining intrinsic GOPs that are invariant to rigid transformations (no need for object pre-alignment) which supports statistical analysis. We will discuss how we can fit a new type of skeletal representation, namely the discrete swept skeletal representation, to a globular object. Finally, we will study the introduced shape representation based on simulated data and data from the ParkWest study. We compare the shapes of the left hippocampi of patients with Parkinson's disease versus a healthy control group.

CO101 Room 45 VARIABLE SELECTION, MODEL SELECTION AND NONPARAMETRIC METHODS Chair: Marialuisa Restaino

C0200: **High-dimensional variable selection in the presence of missing data**
Presenter: **Philip Yu**, The Education University of Hong Kong, Hong Kong
Co-authors: Lixing Liang, Yipeng Zhuang

Regression analysis is often affected by high dimensionality, severe multicollinearity, and a large proportion of missing data. These problems may mask important relationships and even lead to biased conclusions. A novel computationally efficient method is proposed that integrates data imputation and variable selection to address these issues. More specifically, the proposed method incorporates a new multiple imputation algorithm based on matrix completion (Multiple Accelerated Inexact Soft-Impute), a more stable and accurate new randomized lasso method (Hybrid Random Lasso), and a consistent method to integrate a variable selection method with multiple imputation. Compared to existing methodologies, the proposed approach offers greater accuracy and consistency through mechanisms that enhance robustness against different missing data patterns and sampling variations. The method is applied to analyze the Asian American minority subgroup in the 2017 National Youth Risk Behavior Survey, where key risk factors related to the intention for suicide among Asian Americans are studied. The proposed method demonstrates enhanced accuracy, consistency, and efficiency in variable selection and prediction through simulations and real data analyses in various regression and classification settings.

C0254: **BRBVS: variable ranking in copula survival models affected by general censoring scheme**
Presenter: **Danilo Petti**, University of Essex, United Kingdom
Co-authors: Marcella Niglio, Marialuisa Restaino

The newly developed BRBVS package is introduced and discussed. BRBVS presents an innovative approach to variable selection in the presence of bivariate time-to-event data, characterized by censoring/truncation and correlation. This tool allows researchers to identify two sets of relevant covariates by a new metric that considers the dependency structure between survival functions. The effectiveness of BRBVS is demonstrated through numerical and graphical results from an extensive simulation study and with the analysis of a data set collected from a study on age-related eye disease.

C0304: **A boosting method for variance components selection in linear mixed models**
Presenter: **Michela Battauz**, University of Udine, Italy
Co-authors: Paolo Vidoni

Boosting is a method developed in machine learning and later translated to statistical analysis to estimate the parameters of a model. The procedure iteratively improves the fit of the model by updating a subset of parameters at each step. In this approach, early stopping is fundamental to perform model selection and prevent overfitting. The proposals in the literature for random effects models focus on the fixed part of the model, while the variables with random effects should be pre-specified. We present a novel method to select the variance components of the model that considers the negative profile log-likelihood as the objective function to minimize. The issue of non-convexity of such a function is overcome by exploiting the directions of negative curvature, which allows scaping saddle points or local maxima. Simulation studies show the good performance of the proposal in detecting the real structure of the data, while an application to the analysis of total nitrate concentration further illustrates the procedure.

C0341: **New software developments for the analysis of ranking data**
Presenter: **Michael Georg Schimek**, Medical University of Graz, Austria

In various fields of application, lists of distinct objects are presented in rank order because one can always rank objects according to their position on a scale. An observed ordering might be due to a measure of the strength of evidence, an assessment based on expert knowledge, or a technical device. Also, variable values can be replaced by corresponding ranks, but the resulting loss of accuracy is compensated by a gain in generality. The fact that rankings are invariant under the stretching of the scale is a major advantage of this kind of data representation. Various statistical tasks can be performed: (i) measuring the association between ranked lists, (ii) measuring the distance between ranked lists, (iii) identification of significantly overlapping sublists (estimation of the point of degeneration of paired rankings into noise), (iv) aggregation (consolidation) of ranked lists or sublist, and (v) reconstruction of the signals (i.e. estimation of latent parameters) that inform observed ranked lists. The use of the recent R packages on CRAN, TopKLists and TopKSignal, developed by the author and co-workers, is exemplified for selected tasks.

C0354: **Robust inference for the unification of confidence intervals in meta-analysis**
Presenter: **Hongsheng Dai**, Newcastle University, United Kingdom
Co-authors: Wei Liang, Yinghui Wei, Haicheng Huang

Traditional meta-analysis assumes that the effect sizes estimated in individual studies follow a Gaussian distribution. However, this distributional assumption is not always satisfied in practice, leading to potentially biased results. In a situation where the number of studies is large, the cumulative Gaussian approximation errors from each study could render the final estimation unreliable. An empirical likelihood method is developed for combining confidence intervals under the meta-analysis framework. This method is free of the Gaussian assumption in effect size estimates from individual studies and from the random effects. This new methodology supersedes conventional meta-analysis techniques in both theoretical robustness and computational efficiency.

CC027 Room 001 MACHINE LEARNING Chair: Emese Lazar

C0259: **Even naive trees are consistent**
Presenter: **Nico Foege**, Otto-von-Guericke University Magdeburg, Germany
Co-authors: Markus Pauly, Lena Schmid, Marc Ditzhaus

Tree-based methods such as Random Forests are learning algorithms that have become an integral part of the statistical toolbox. The last decade has shed some light on theoretical properties, such as their consistency for regression tasks. We illustrate that consistency results are, in general, not enough. To this end, we introduce a new class of naive trees, which do the subspacing completely at random and independent of the data. We then provide direct proof of their consistency. Since naive trees appear to be too simple for actual application, we further analyze their finite sample properties in a simulation and small benchmark study. We find a slow convergence speed and a rather poor predictive performance. Based on these results, we finally discuss to what extent consistency proofs help to justify the application of complex learning algorithms.

C0405: **The hidden algebra of interactive visualization: Exploring the links between graphics, statistics, and interaction**
Presenter: **Adam Bartonicek**, The University of Auckland, New Zealand

With the rise of web technologies, interactive data visualizations have become a staple of data presentation. Yet, despite the growing popularity of interactive graphics, a formal framework is still lacking for turning raw data into statistical summaries that can be interactively visualized. The reason for this lack may be a subtle yet profound issue: while it is often desirable to treat the statistical summaries and geometric objects in the plots as independent components, this is rarely the case. Consider an ordinary stacked barplot. Data visualization researchers have long warned that while stacking some statistics, such as counts or sums, will produce a valid overall statistic, stacking others will not. But what are the mathematical properties that make this possible? Can other summaries be stacked? And why do stacked plots, despite general reluctance in static visualization, still enjoy wide popularity in interactive visualization? The purpose is to delve into the relationship between graphics, statistics and interaction. Specifically, by discussing various concepts from category theory, such as monoids and groups, the hope is to provide a new appreciation of the rich structure that lies beyond the figures observed daily. Finally, a new R package, plotscaper, is briefly introduced for interactive data exploration, which is an attempt to implement some of these ideas.

C0407: **Enhancing geometrically designed spline regression through generalized additive models and functional gradient boosting**
Presenter: **Emilio Luis Saenz Guillen**, Bayes Business School, United Kingdom
Co-authors: Dimitrina Dimitrova, Vladimir Kaishev

Geometrically designed spline (GeDS) regression offers an accurate and efficient solution to the spline regression problem by automatically estimating the number and positions of the knots using a stopping rule controlled by two tuning parameters. Two ground-breaking enhancements to the GeDS methodology are introduced. First, the applicability of GeDS is broadened to cover the family of generalized additive models (GAM) by implementing the local-scoring algorithm using GeD splines as function smoothers. Second, functional gradient boosting (FGB) is integrated to dynamically estimate the number and locations of the knots, as well as the regression coefficients of the spline model. Unlike typical gradient boosting models that generally lack an interpretable representation, the final FGB-GeDS fit is expressed as a single spline model. Additionally, FGB-GeDS automatically determines two main boosting parameters: the number of boosting iterations and the shrinkage/learning rate. On the one hand, the number of boosting iterations is regulated by a stopping rule analogous to the one used in the canonical GeDS method. On the other hand, the weakness of the GeDS base learners is controlled internally by the tuning parameters of the GeDS stopping rule, thus eliminating the need for additional regularization parameters like a shrinkage/learning rate.

C0423: **Multi-attribute preferences: A transfer learning approach**
Presenter: **Sjoerd Hermes**, Wageningen University, Netherlands
Co-authors: Joost van Heerwaarden, Pariya Behrouzi

A novel statistical learning methodology is introduced based on the Bradley-Terry method for pairwise comparisons, where the novelty arises from the method's capacity to estimate the worth of objects for a primary attribute by incorporating data of secondary attributes. These attributes are properties on which individuals evaluate objects in a pairwise fashion. By assuming that the main interest of practitioners lies in the primary attribute and that the secondary attributes only serve to improve the estimation of the parameters underlying the primary attribute, the well-known transfer learning framework is utilised. To wit, the proposed method first estimates a biased worth vector using data pertaining to both the primary attribute and the set of informative secondary attributes, which is followed by a debiasing step based on a penalized likelihood of the primary attribute. When the set of informative secondary attributes is unknown, we allow for their estimation by a data-driven algorithm. Theoretically, it is shown that, under mild conditions, the ℓ_∞ and ℓ_2 rates are improved compared to fitting a Bradley-Terry model on just the data pertaining to the primary attribute. The favorable (comparative) performance under more general settings is shown by means of a simulation study.

C0503: **Bayesian federated inference for estimating statistical models based on non-shared multicenter data sets**

Presenter: **Marianne Jonker**, Radboud university medical center, Netherlands

Co-authors: Hassan Pazira, Emanuele Massa, Ton Coolen

Identifying predictive factors via multivariable statistical analysis is often impossible for rare diseases because the available data sets are too small. Combining data from different medical centers into a single (larger) database would alleviate this problem but it is, in practice, challenging due to regulatory and logistic problems. A Bayesian federated inference (BFI) framework is proposed. It aims to construct from local inferences in separate data centers what would have been inferred had the data sets been merged. It can cope with small data sets by inferring locally not only the optimal parameter values but also additional features of the posterior parameter distribution. The BFI methodology has been developed for generalized linear models and survival models, for homogeneous and heterogeneous populations, and for association and prediction models. The performance of the proposed methodology is quantified using simulated and real-life data.

CC058 Room 051 COMPUTATIONAL STATISTICS Chair: Stefanie Biedermann

C0398: **Derivation of optimal solution by full enumeration for subgroup identification**

Presenter: **Masahiro Mizuta**, The Institute of Statistical Mathematics, Japan

Subgroup identification methods are techniques for identifying subsets for which a particular treatment, etc., is effective. For example, a widely used expression in relation to COVID-19 is "recommend vaccination for people of a certain age". Identifying valid subsets is a challenging task, and several approaches have been proposed (Pmtree, glmtree, QUINT, etc.). However, these methods do not guarantee that the optimal subgroup with the maximum validity evaluation function is obtained. Consequently, techniques that determine the "optimal subgroups in the strict sense" by enumerating all subgroups and computing the effectiveness of each are effective. For instance, the algorithm that is developed enumerates 43,199,128,758 (43.2 billion) subgroups in artificial data with p=5 and n=100. The optimal subgroups can then be derived. Furthermore, realistic modifications of the algorithm have also been implemented. Subgroup identification is the foundation of subgroup analysis and is a crucial technique in personalized medicine. On the other hand, the treatment of subgroups must be meticulously considered when implementing statistical decisions. The findings can be utilized to contribute to subgroup identification and analysis.

C0472: **Categorical encoding as joint optimization in predictive models**

Presenter: **Iris-Ioana Roatis**, Imperial College London, United Kingdom

Co-authors: Ed Cohen, Niall Adams

The necessity of handling categorical variables, which are not inherently numerical, is a significant challenge in predictive modelling. Developing efficient methods to encode these variables, particularly those with high cardinality, is crucial. While the literature conceptualises the prediction process as comprising two distinct stages, encoding followed by model training, a novel approach is proposed. The new idea consists of jointly optimising the two steps and hence treating it as one single task. This method preserves model interpretability with the advantage of eliminating the need to choose among existing encoding techniques. The embedding is viewed as a non-linear combination of the chosen characteristics of the data. For example, for binary classification problems, the counts of positive and negative labels within each category are considered, while for regression problems, the average and variance of all entries within that category are used. The resulting numerical representation and the remaining features are used to train the model for predicting the target variable, with the loss being backpropagated to jointly update the embedding of the categorical variables. The behavior of this proposal is demonstrated through a series of experiments on simulated and real-life data with promising outcomes.

C0158: **Constructing Bayesian optimal designs for discrete choice experiments by simulated annealing**

Presenter: **Yicheng Mao**, Maastricht University, Netherlands

Co-authors: Roselinde Kessels, Tom van der Zanden

Discrete Choice Experiments (DCEs) investigate the attributes that influence individuals' choices when selecting among various options and are widely applied across numerous fields. To enhance the quality of the estimated choice models, many researchers opt for Bayesian optimal designs that take into account already existing information about the attributes' preferences. Given the nonlinear nature of choice models, the construction of an appropriate design necessitates the use of efficient algorithms. Among these, the Coordinate-Exchange (CE) algorithm is most commonly employed for constructing designs based on the multinomial logit model. This algorithm cannot guarantee globally optimal designs, and obtaining better designs often requires the use of multiple independent random starting designs, significantly increasing the algorithm's computational load. We propose the use of Simulated Annealing (SA) to construct Bayesian D-optimal designs. The SA algorithm does not require the use of multiple random starting designs, offering greater computational efficacy than the CE algorithm. Our work represents the first application of the SA algorithm in constructing Bayesian optimal designs for DCEs. Through multiple computational experiments and a real-life case study, we compare the performance of the algorithms, finding that the SA designs consistently outperform the CE designs in terms of Bayesian D-efficiency, especially when the prior preference information is highly uncertain.

C0451: **A penalized maximum likelihood estimation for hidden Markov models to address latent state separation**

Presenter: **Luca Brusa**, University of Milano-Bicocca, Italy

Co-authors: Francesco Bartolucci, Fulvia Pennoni, Romina Peruilh Bagolini

In analyzing longitudinal data, the focus is usually on the evolution of a characteristic of interest over time, which is measured by occasion-specific response variables. To analyze such data, the hidden Markov model assumes a latent process typically following a Markov chain of the first order and affecting the distribution of the response variables. It may include both time-constant and time-varying covariates, which can affect the conditional distribution of the response variable (measurement model). The latent process accounts for unobserved heterogeneity when covariates cannot fully explain the variability among responses. It also allows the consideration of state dependence by including the lagged responses among the covariates. When these covariates do not fully explain the heterogeneity between individuals, the parameters corresponding to the effect of the latent states may be very large, leading to widely separated states that cause instability of the estimated parameters. A penalized likelihood approach is implemented by modifying the M-step of the expectation-maximization algorithm. In addition, a cross-validation method is proposed to jointly select the number of latent states and the strength of the penalty term. The asymptotic properties of the estimator are examined via simulation, and the proposal is illustrated to estimate the effect of covariates on hypotension occurrence during anesthesia performed before surgery.

C0455: **Taming numerical imprecision by adapting the KL divergence to negative probabilities**

Presenter: **Simon Pfahler**, University of Regensburg, Germany

Co-authors: Peter Georg, Rudolf Schill, Maren Klever, Lars Grasedyck, Rainer Spang, Tilo Wettig

The Kullback-Leibler (KL) divergence is frequently used in data science to compare probability distributions. When considering discrete probability vectors on exponentially large state spaces, one typically needs approximations to keep calculations tractable. This may result in approximations of probability vectors with a few small negative entries, rendering the KL divergence undefined. To address this problem, a parameterized substitute divergence measure, the shifted KL (sKL) divergence, is introduced. In contrast to existing techniques, the approach is not problem-specific and does not increase the computational overhead. The sKL divergence retains many of the useful properties of the KL divergence while being resilient to Gaussian noise in the probability vectors. For a large class of parameter choices, it is proven that the sKL divergence converges to the KL divergence in the limit of small Gaussian noise. In a concrete example that does not satisfy the assumption of Gaussian noise, the tensor-train approximation, it is shown that the method still works reliably. As an example application, it is also shown how the approach can be used in bioinformatics to accelerate the optimization of mutual hazard networks, a type of cancer-progression model.

CC030 Room 052 FORECASTING Chair: Thomas Fung

C0172: **Bregman model averaging for forecast combination**
Presenter: **Chu-An Liu**, Academia Sinica, Taiwan
Co-authors: Yi-Ting Chen, Jiun-Hua Su

A unified model averaging (MA) approach is provided, and its asymptotic optimality is established for a wide class of forecasting targets. The asymptotic optimality is achieved by minimizing an asymptotic risk based on the expected Bregman divergence of a combined-forecast sequence from a forecasting-target sequence under the local(-to-zero) asymptotics. This approach is flexibly applicable to generate MA methods in different forecasting contexts, including, but not limited to, univariate or multivariate mean forecasts, volatility forecasts, probabilistic forecasts and density forecasts. We also conduct Monte Carlo simulations (empirical applications) to show that compared to related existing methods, the MA methods generated by this approach perform reasonably well in finite samples (real data).

C0178: **Forecast combination and interpretability using random subspace**
Presenter: **Boris Kozyrev**, Halle Institute for Economic Research (IWH), Germany

Forecast aggregation is investigated via random subspace regressions (RS), and the potential link between RS and the Shapley value decomposition (SVD) is explored using the US GDP growth rates. This combination of techniques enables handling high-dimensional data and reveals the relative importance of each individual forecast. First, we demonstrate that in certain practical instances, it is possible to enhance forecasting performance by randomly selecting smaller subsets of individual forecasts and obtaining a new set of predictions based on a regression-based weighting scheme. The optimal value of selected individual forecasts is also empirically studied. Then, we propose a connection between RS and the SVD, enabling the examination of each individual forecast's contribution to the final prediction, even when the number of forecasts is relatively large. This approach is model-agnostic (can be applied to any set of forecasts) and facilitates understanding of how the aggregated prediction is obtained based on individual forecasts, which is crucial for decision-makers.

C0293: **Boosting XGBoost: Using the panel dimension to improve machine-learning-based forecasts in macroeconomics**
Presenter: **Johannes Frank**, Friedrich-Alexander University Erlangen/Nuremberg, Germany
Co-authors: Jonas Dovern

The short-time dimension of commonly used macroeconomic data sets presents challenges for the estimation of machine learning models designed for real-time business cycle monitoring. We consider panel data to increase the data set available for training and nowcasting US unemployment using extreme gradient boosting and neural networks. The underlying idea is that dynamics between variables and across time at the state level are similar to each other and to the dynamics at the national level. We use data pooling in combination with weight sharing that accommodates some cross-sectional heterogeneity. This approach facilitates parameter regularization and safeguards against overfitting. We find that this soft pooling approach improves forecast accuracy at the national level and reduces both the variance and the mean of the RMSE distribution across states. Thus, leveraging regional information in a panel data framework with suitable regularization techniques addresses data scarcity in macroeconomic nowcasting effectively.

C0389: **Univariate time series forecasting using echo state networks: An empirical application**
Presenter: **Alexander Haeusser**, Justus-Liebrig-University Giessen, Germany

The echo state network (ESN) is introduced for univariate time series forecasting. The echo state approach incorporates building a large dynamic reservoir, which models non-linear relationships between inputs and outputs to capture time series patterns like autocorrelation, trend, and seasonality. Even when the model is non-linear, training of the ESN is simplified to a linear model, estimated via Ridge Regression. A standard procedure for fast model estimation and selection is introduced. The modelling framework is suited for a wide range of applications, and an empirical analysis of real-world data from the M4 Forecasting Competition illustrates its modelling flexibility and forecast accuracy. The proposed echo state network can outperform or compete against state-of-the-art forecasting models from statistics and machine learning.

C0418: **Enhancing electricity price forecasting accuracy: A novel filtering strategy for improved out-of-sample predictions**
Presenter: **Andrea Cerasa**, European Commission - Joint Research Centre, Italy
Co-authors: Alessandro Zani

Electricity price forecasting (EPF) has become a crucial component in energy companies' operational strategy. An original filtering strategy is introduced, aimed at refining the accuracy of day-ahead EPF where outliers identification and replacement rely on a model-based procedure. Extreme spikes are identified through the standardized residuals from rolling window robust regressions of prices against a predefined set of regressors. They are then replaced by the values fitted by the model. This method offers several benefits, such as eliminating the need for prior decomposition of price series, reducing the number of choices typical of standard filtering procedures for outlier identification and replacement, and mitigating the issues of masking and swamping by robust methods. The filtering strategy is applied to open-access benchmark datasets of 5-day-ahead markets, using state-of-the-art models and accuracy metrics, and compared to a baseline no-filtering strategy. Empirical results demonstrate that the proposed filtering approach can significantly enhance the precision of EPF while maintaining reasonable computation times. The proposed method offers an efficient pre-processing tool that, through more accurate price forecasts, can significantly improve the optimization of operational and strategic decision-making in the energy sector. It can valuably support energy traders, companies, and generators in mitigating risks and enhancing profitability in day-ahead markets.

CC046 Room 44 TEXT MINING Chair: Francesco Audrino

C0285: **Seeded Poisson factorization: Leveraging domain knowledge to fit topic models**
Presenter: **Bernd Prostmaier**, Paris-Lodron-University Salzburg, Austria
Co-authors: Bettina Gruen, Paul Hofmarcher

The latent variable model Seeded Poisson Factorization (SPF) is proposed, which addresses the challenges in text classification where no labelled texts are available, but the classes are characterized with a set of relevant words. In various business contexts, including, in particular, the assessment of consumer feedback, vast amounts of unlabeled text data are collected where conceptual frameworks outline potential categorization schemata, and domain experts are able to provide sets of relevant words for each category. SPF builds on the Poisson Factorization topic model, which assumes that term counts in documents are independently drawn from a Poisson distribution with the rate resulting from a combination of topic-specific term distributions weighted by the document-specific topic distributions. Seeding modifies the prior distribution of the topic-specific term distributions with the set of relevant words a-priori having higher rates for their topic. Estimation is based on computationally efficient variational inference using general-purpose stochastic gradient optimization. The use of SPF is illustrated on Amazon customer feedback data to classify feedback items where the categories are a-priori known. Empirical results indicate that SPF surpasses alternative topic models, allowing for the specification of seed words for topics in terms of computational cost and classification accuracy.

C0426: **LongFinBERT: A language model for long financial documents**
Presenter: **Erik-Jan Senn**, University of St. Gallen, Switzerland
Co-authors: Minh Tri Phan

LongFinBERT, a modern language model specialized for processing long financial documents, is introduced. Due to an adaptation in model architecture, LongFinBERT demonstrates substantially lower computational requirements for lengthy documents compared to other state-of-the-art language models. This characteristic enables the processing of, e.g. an entire annual accounting filing at once, which was previously computationally infeasible for LMs. LongFinBERT is applied to two empirical settings. Firstly, the aim is to improve the detection of financial misreporting using text from 10-K filings from 1994 to 2018. Misreporting predictions that utilize text-based features from LongFinBERT outperform those based solely on accounting variables or other textual models, namely latent Dirichlet allocation, neural document embeddings, and FinBERT. Lastly, it is found that market returns respond to year-over-year alterations of accounting disclosures, measured using LongFinBERT.

C0436: **Does sentiment help in asset pricing? A novel approach using large language models and market-based labels**
Presenter: **Jule Schuettler**, University of St.Gallen, Switzerland
Co-authors: Francesco Audrino, Fabio Sigrist

A novel approach to sentiment analysis in financial markets is presented that addresses existing challenges in NLP by deriving labels in a data-driven manner. The study is based on an extensive dataset containing financial text from earnings call transcripts, headlines, and tweets for all stocks that belong to the CRSP universe. SMARTyBERT (Sentiment Model with Additional Regressor for Text type Bidirectional Encoder Representation for Transformers) is implemented, which is a pre-trained DeBERTa model expanded with an additional feature indicating the text source. This model is fine-tuned on the specific dataset, harnessing the power of transfer learning to capture the relationship between investor sentiment and next-day excess returns. Empirical asset pricing demonstrates the predictive power of sentiment extracted by SMARTyBERT, with an equal-weighted long-short strategy yielding an annualized mean return of 45% and a Sharpe ratio of 2.78. The comparison with FinBERT underscores the superiority of the data-driven labeling approach over traditional human-annotated labeling.

C0448: **Variational inference for the keyword assisted topic models**
Presenter: **Kiyoshi Inoue**, Doshisha University Graduate School, Japan
Co-authors: Shintaro Yuki, Yoshikazu Terada, Hiroshi Yadohisa

Latent Dirichlet allocation (LDA) is often applied to discover latent topics in documents. Sometimes, some keywords for some topics are known in advance. As an extension of LDA, the keyword assisted topic model (KeyATM) has been proposed to improve the quality of the estimated topic and to provide more interpretable results by incorporating keywords for each topic as prior information. The KeyATM uses the collapsed Gibbs sampling for estimation. However, it is known that the Gibbs sampler is slow, and thus, the current algorithm of the KeyATM is not scalable for large-scale data. Therefore, a variational inference algorithm is developed for the KeyATM, which is faster than the Gibbs sampling algorithm. The proposed algorithm of the KeyATM can be applied to large-scale data. The advantages of the proposed algorithm are validated through numerical experiments and real data application.

C0452: **Text data insights and machine learning innovations in monetary policy shock identification**
Presenter: **Nickson Cabote**, Washington State University, United States

A new method is introduced to identify monetary policy shocks in the Philippines and enhance macroeconomic forecasts by integrating textual data from the Philippine Central Bank's policy meeting records. Utilizing machine learning techniques and Natural Language Processing (NLP), such as sentiment analysis and TF-IDF, this approach extends previous frameworks. It analyzes exogenous variations in monetary rates independently of central bank macroeconomic forecasts. The integration of Principal Component Analysis (PCA) of macroeconomic factors with textual data significantly enhances forecast accuracy, improving predictions for GDP growth and inflation by up to 42% and 8%, respectively, compared to traditional factor-augmented VAR models. Moreover, the refined models, which incorporate a comprehensive set of textual and economic indicators, explain up to 89% of the variance in policy rates from 2002 to 2020. Applying these monetary policy shocks with the local projections method reveals a quicker GDP response and addresses the 'price puzzle' commonly seen in standard VARs in developing economies. This approach underscores the potential of advanced word vector models to deepen the understanding of monetary policy dynamics in emerging markets.

Thursday 29.08.2024 09:00 - 10:00 Parallel Session H – COMPSTAT2024

CV040 Room 051 BIOSTATISTICS Chair: Andreas Artemiou

C0487: **K-cover: A novel way to compare distributions in the context of drug development**
Presenter: **Gerhard Goessler**, University of Graz, Austria
Co-authors: Vera Hofer, Hans Manner, Walter Goessler

One of the hurdles in the development of generic drugs is frequently the statistical demonstration of the similarity of the test product (generic) to a reference medicinal product (originator) concerning several quality attributes (QAs). It raises the question of how to describe similarity in a suitable mathematical way. A mathematical concept is proposed for comparing the QA distributions that take account of the asymmetrical nature of the problem, i.e. the distribution of the QA of the test product has to be suitably covered by that of the reference product. For the QA distribution of the test product, this approach allows the central fraction of the probability mass to vary within certain limits derived from the QA distribution of the reference product while simultaneously demanding strict restrictions on the more extreme parts, i.e., the tails, of the test product distribution. It allows for the account of consumer safety in situations where no information is available on specified values and tolerances.

C0488: **A statistical test for the overlap of normal distributions based on generalized p-values**
Presenter: **Vera Hofer**, University of Graz, Austria
Co-authors: Gerhard Goessler, Hans Manner, Walter Goessler

In the context of the development of generic drugs, it is often necessary to test whether the distribution of a quality attribute (QA) of a test product (generic drug) is suitably covered by a reference product. Based on k-cover, a statistical test for simultaneously comparing the k- and the (1-k)-quantiles of the test and the reference distribution is proposed, assuming both distributions are normal. It constitutes a multiple testing problem that demands, depending on the distributions at hand, a proper correction (similar to Bonferroni correction) to guarantee that the significance level chosen is maintained. The proposed test is an extension of an existing test for comparing a quantile of two distributions, which utilizes the theory of generalized inference. It results in a two-step test procedure, which implicitly performs a pre-test that decides the correction to apply to avoid error inflation due to multiple testing. Simulations show that the test proposed keeps the type-I error under control and has the desired asymptotic properties.

C0412: **Outlier detection in mass-spectrometry data using the conformal prediction framework**
Presenter: **Soohyun Ahn**, Ajou University, Korea, South

Quality control procedures are crucial for ensuring the reliability of mass spectrometry (MS) data, which is vital in biomarker discovery and understanding complex biological systems. However, existing methods often concentrate solely on either sample or peak outlier detection, relying on subjective criteria or employing overly uniform thresholds based on asymptotic distributions, thus failing to adequately reflect the characteristics of the data. A novel approach leveraging conformal prediction for outlier detection is introduced in MS data analysis. The method simultaneously identifies outlier samples and peaks based on data-driven and distribution-free principles. Rigorous numerical evaluations and comparisons with an existing method demonstrate superior diagnostic performance. Application to real LC-MRM data underscores practical utility, enhancing data reliability and reproducibility in MS studies.

CO125 Room 44 HITEC: COMPLEX DATA Chair: Bojana Milosevic

C0395: **Clustering in multiway networks**
Presenter: **Vladimir Batagelj**, IMFM, Slovenia

A weighted multiway network $\mathbf{N} = (\mathbf{V}, \mathbf{L}, w)$ is based on nodes from k finite sets (ways or dimensions) $\mathbf{V} = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_k)$, the set of links $\mathbf{L}$, and the weight $w : \mathbf{L} \to R$. The incidence function $I : \mathbf{L} \to \mathbf{V}_1 \times \mathbf{V}_2 \times \cdots \times \mathbf{V}_k$ assigns to each link $e \in \mathbf{L}$ a k-tuple of its nodes $I(e) = (e(1), e(2), \ldots, e(i), \ldots, e(k))$, $e(i) \in \mathbf{V}_i$. If for $i \neq j$, $\mathbf{V}_i = \mathbf{V}_j$, we say that $\mathbf{V}_i$ and $\mathbf{V}_j$ are of the same mode. In a general multiway network, different additional data can be known for nodes and/or links $\mathbf{N} = (\mathbf{V}, \mathbf{L}, \mathbf{P}, \mathbf{W})$, where $\mathbf{P}$ is a set of node properties $p : \mathbf{V}_i \to S_p$, and $\mathbf{W}$ is a set of link weights $w : \mathbf{L} \to S_w$. An approach to clustering in multiway networks is discussed. Extending the projection approach from two-mode networks, some options are first explored to define the projection in a selected way. The obtained projection matrices can be transformed in the generalized Salton and Jaccard similarity measures that can be used for clustering ways using standard clustering procedures. The proposed approach is illustrated by analyzing selected data from ESS - European Social Survey 2023. The approach is supported by the R package MWnets.

C0489: **On the application of kernel-based independence tests to variable selection problems**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Jelena Radojevic

Kernel-based generalizations of distance covariance are explored and are applied to variable screening procedures. The flexibility of this association measure allows for the inclusion of models with spherical and hyperspherical data, which are common in various applied research fields such as meteorology, geology, biology, and more. The robustness and adaptability of the proposed method are demonstrated through extensive empirical studies. Overall, the findings suggest that kernel-based distance covariance is a powerful tool for variable selection in high-dimensional datasets.

C0495: **Enhancing subarachnoid hemorrhage monitoring with AI and uncertainty analysis**
Presenter: **Robertas Alzbutas**, Kaunas University of Technology, Lithuania
Co-authors: Tomas Iesmantas, Jewel Sengupta

The focus is on improving the monitoring of subarachnoid hemorrhage (SAH) through advanced artificial intelligence techniques. Traditional segmentation models are enhanced by integrating image augmentation and adding a regression layer, enabling more precise localization and quantification of SAH from cerebral images. Additionally, a modified region-growing method segments affected brain regions, from which features are extracted using pre-trained models. Dimensionality reduction is applied via optimization algorithms, and an unsupervised deep learning model based on spatial distance is used for automatic SAH segmentation. This model achieves accurate, regular contours quickly. Additionally, optimized feature vectors are classified using an autoencoder to identify SAH subtypes. The approach's effectiveness is validated through case studies, emphasizing feature selection and prediction accuracy, aiding clinicians in making informed treatment decisions. The presented framework integrates AI with uncertainty analysis, combining multiple models to assess prediction performance, analyze result sensitivity, and reduce overall uncertainty through sampling-based methods and tolerance intervals. This methodology facilitates timely data incorporation and improved training processes, enhancing the identification of critical contributors to such SAH monitoring.

CO094 Room 45 RESAMPLING METHODS FOR DEPENDENT DATA Chair: Claudia Kirch

C0255: **Optimal choice of bootstrap block length for periodically correlated time series**
Presenter: **Anna Dudek**, AGH University of Krakow, Poland

The focus is on the problem of choosing the optimal block length for block bootstrap methods designed for periodically correlated processes, such as the Generalized Seasonal Block Bootstrap, the Extension of Moving Block Bootstrap, and the Generalized Seasonal Tapered Block Bootstrap.

We consider two estimation problems: the overall mean and the seasonal means. The obtained optimal block lengths are of the same order as for the corresponding approaches in the stationary case. They are obtained directly by minimizing the mean squared error of the corresponding bootstrap variance estimator or by exploiting the relationship between bootstrap and jackknife variance estimators. Finally, we present the results of the performed simulation study.

C0291: **A bootstrap-assisted self-normalization approach to inference in cointegrating regressions**
Presenter: **Carsten Jentsch**, TU Dortmund University, Germany
Co-authors: Karsten Reichold

Traditional inference in cointegrating regressions requires tuning parameter choices to estimate a long-run variance parameter. Even in case these choices are "optimal", the tests are severely size distorted. We propose a novel self-normalization approach, which leads to a nuisance parameter-free limiting distribution without estimating the long-run variance parameter directly. This makes our self-normalized test tuning parameter-free and considerably less prone to size distortions at the cost of only small power losses. In combination with an asymptotically justified vector autoregressive sieve bootstrap to construct critical values, the self-normalization approach shows further improvement in small to medium samples when the level of error serial correlation or regressor endogeneity is large. We illustrate the usefulness of the bootstrap-assisted self-normalized test in empirical applications by analyzing the validity of the Fisher effect in Germany and the United States.

C0314: **A practical resampling-based approach to interval estimation for spectral densities**
Presenter: **Daniel Nordman**, Iowa State University, United States

The spectral density function can play an important role in time series analysis, where nonparametric interval estimation of the spectral density becomes useful. Existing interval methods for spectral densities, including chi-square approximation and frequency domain bootstrap (FDB), can often be problematic in practice because intervals exhibit poor coverage. An alternative approach is presented that merges empirical likelihood (EL) and FDB. The idea is that EL provides a new statistic for setting intervals for spectral densities that can be effectively calibrated by bootstrap. The FDB-EL procedure is valid under mild conditions for application to a wide range of time processes. Numerical studies suggest that FDB-EL confidence intervals exhibit good performance compared to other methods. The confidence interval procedure is illustrated with an application to wind turbines.

CC130 Room 001 EXTREME VALUES Chair: Jean Marc Bardet

C0165: **Generalized random forest for extreme quantile regression**
Presenter: **Mahutin Lucien Vidagbandji**, LMAH-University of Le Havre Normandy, France
Co-authors: Alexandre Berred, Cyrille Bertelle, Laurent Amanton

Quantile regression is a commonly used statistical method in regression analysis. In contrast to classical regression, which centers on predicting the conditional mean of a dependent variable based on independent variables, quantile regression aims to predict conditional quantiles. Specifically, if $Y \in \mathcal{Y} \subset \mathbb{R}$ represents a random variable describing a risk factor dependent on a set of covariates represented by the random vector $X \in \mathcal{X} \subset \mathbb{R}^p$, the goal is to estimate the conditional extreme quantile given by: $Q_\tau(x) = \inf\{y : F^{-1}_{Y|X=x}(y) \geq \tau\}$ with $\tau \in [0,1]$. Classical quantile regression methods face challenges, especially when the quantile of interest is extreme, due to the limited number of data available in the tail of the distribution or when the quantile function is complex. We propose an extreme quantile regression method based on extreme value theory and statistical learning to overcome these challenges. Following the Block Maxima (BM) approach of extreme value theory, we approximate the conditional distribution of BM by the generalized extreme value distribution, with parameters depending on covariates. To estimate these parameters, we employ a method based on generalized random forests. Simulated data applications highlight our method's performance compared to other statistical learning-based quantile regression approaches.

C0358: **Modeling waiting times of clustered extreme events with application to mid-latitude winter cyclones**
Presenter: **Christina Meschede**, TU Dortmund University, Germany
Co-authors: Katharina Hees, Roland Fried

For many applications in the field of extreme value theory, both the frequency of the occurrence and the return times of extreme events are of great interest, such as in climate research in the study of extreme mid-latitude cyclones. Traditionally, a Poisson process is assumed as a model for the occurrence of extreme events so that the waiting times between two successive exceedances are i.i.d. exponentially distributed. However, this does not properly reflect the temporal clustering that often occurs in data. A prior study provided a modelling framework for such clustering behaviour under the assumption that the observations are realizations of a strictly stationary process with an existing extremal index. In such cases, the scaled waiting times between extreme events are approximately distributed as a mixture of an exponential distribution and a Dirac measure of zero. A recent study proposed another model for clustered extreme events based on a fractional Poisson process, leading to Mittag-Leffler distributed inter-exceedance times. The purpose is to introduce a generalized model that includes exponential, mixed, and Mittag-Leffler distributed waiting times as special cases. The suitability of a minimum distance method is verified based on a modification of the Cramr-von Mises distance for joint estimation of the model parameters and an application on climate data is shown.

C0220: **Comparative study on tail probability estimators**
Presenter: **Taku Moriyama**, Yokohama City University, Japan

Tail probability estimators in iid settings are considered. There are mainly two ways for the estimation; the fitting to the generalized Pareto distribution and the fully nonparametric estimation. The fitting estimator is justified by the approximation proven in the extreme value theory; however, the accuracy depends on how extremely large the target is. The nonparametric estimator does not need the approximation and has the advantage of wide applicability. Both theoretical and numerical comparative studies on excess distribution estimation are conducted. Asymptotic convergence rates of estimators are obtained, and the mean integrated squared errors are numerically surveyed by simulation study.

CC136 Room 050 STATISTICAL MODELS AND INFERENCE FOR APPLICATIONS Chair: Thomas Fung

C0177: **Inference for multicomponent stress-strength reliability based on generalized Lindley distribution**
Presenter: **Fatma Gul Akgul**, Karadeniz Technical University, Turkey

The classical and Bayesian estimation of reliability in the multicomponent stress strength model when both the stress and strengths are drawn from the generalized Lindley distribution are considered. The maximum likelihood (ML) and Bayesian methods are used in the estimation procedure. The Bayesian estimates of reliability are obtained by using Lindley's approximation, Tierney-Kadane approximation and Markov Chain Monte Carlo (MCMC) methods due to the lack of explicit forms. The asymptotic confidence intervals are constructed based on the ML estimators. For small sample sizes, the bootstrap confidence intervals are considered. The MCMC method is used to construct the Bayesian credible intervals. A Monte Carlo simulation study is conducted to investigate and compare the performance of the proposed methods. Finally, a real data set is analyzed for illustrative purposes.

C0176: **Inference and diagnostics for a heteroscedastic partially linear model with skew heavy-tailed error distribution**
Presenter: **Fatma zehra Dogru**, Giresun University, Turkey
Co-authors: Olcay Arslan

Partially Linear Models (PLMs) have gained attention from researchers as valuable tools for dealing with diverse data sets, particularly in fields such as economics and biometrics. Traditionally, PLMs assume a normal distribution for the error term. However, real-world data often deviate from this assumption, exhibiting skewness, heavy-tailed, and heteroscedasticity. Therefore, the Heteroscedastic Partial Linear Model (HPLM) under the Skew Laplace Normal (SLN) distribution is investigated. This model aims to address skewness, heavy-tailing, and heteroscedasticity simultaneously. The model parameters are estimated by Maximum Likelihood Estimation (MLE) using the Expectation/Conditional Maximisation (ECM) algorithm. The influence of diagnostics tailored to the HPLM-SLN model is also examined. In addition, a likelihood ratio test is introduced to assess the homogeneity of the scale parameter, providing insight into the variance homogeneity assumptions. Extensive simulation studies are performed to evaluate the performance of the ECM algorithm and the likelihood ratio test in terms of variance homogeneity. Finally, the effectiveness of the HPLM-SLN model is demonstrated through its application to a real-world dataset on ragweed pollen concentration, demonstrating its utility in practical scenarios.

C0440: **Dynamic linear mixed models for time-dependent data analysis**
Presenter: **Dario Ferreira**, University of Beira Interior, Portugal
Co-authors: Sandra Ferreira, Patricia Antunes, Gilberto Neves

A time-varying linear mixed model (TVLMM), an innovative approach for analyzing time-dependent data, is introduced. Unlike traditional time-series models that assume constant random effects, TVLMM incorporates random effects that change over time, thereby providing a more realistic representation of real-world data dynamics. By integrating the predictive strengths of autoregressive integrated moving average (ARIMA) models with the flexibility of linear mixed models (LMMs), TVLMM addresses the limitations of conventional models in handling temporal variations in data. This approach is particularly relevant for fields such as finance, economics, social sciences, and biology, where the underlying data structures often evolve over time. The methodologies of ARIMA and LMMs are outlined, the parameter estimation process for TVLMM is detailed, and its application is illustrated through a numerical example. The results demonstrate the model's capability to produce accurate predictions by accounting for time-varying characteristics in random effects.

CC131 Room 052 COPULAS IN FINANCIAL ECONOMETRICS Chair: Massimiliano Caporin

C0403: **Dynamic asymmetric tail dependence among multi-asset classes for portfolio management: Dynamic skew-t copula approach**
Presenter: **Toshinao Yoshiba**, Tokyo Metropolitan University, Japan
Co-authors: Kakeru Ito

AC dynamic skew-t copula is proposed with cDCC model to capture dynamic asymmetric tail dependence structure among multi-asset classes (government bonds, corporate bonds, equities, and REITs). The empirical analysis shows that the proposed dynamic AC skew-t copula fits data of multi-asset classes better than other dynamic elliptical copulas, including conventional dynamic skew-t copula, in terms of AIC and BIC. Besides, lower tail dependence coefficients have recently increased compared to upper tail dependence coefficients for all pairs. This indicates that the diversification effect through multi-asset investment has decreased, and investors should enhance tail risk management. Furthermore, out-of-sample analysis shows that using dynamic skew-t copula, especially the proposed model, enhances expected shortfall (ES) estimation accuracy and the performance of minimum ES portfolio compared to dynamic t copula and dynamic normal copula. It indicates that capturing dynamic asymmetric tail dependence is crucial for multi-asset investment.

C0476: **Copula-based clustering of financial time series via evidence accumulation**
Presenter: **Andrea Mecchina**, University of Trieste, Italy
Co-authors: Roberta Pappada, Nicola Torelli

Understanding the dependence structure of asset returns is fundamental in risk assessment and is particularly important in a portfolio diversification strategy. When clustering time series of financial returns, it is largely recognized that pairwise association among values in the left tail of their joint distribution should be considered. To this aim, various solutions using copula models have been proposed to define dissimilarity measures based on finite (lower) tail dependence coefficients. Unfortunately, the result depends on the copula model considered and on some choices attaining the clustering procedures. A clustering approach is proposed where evidence accumulated in a multiplicity of classifications is achieved using classical hierarchical procedures and multiple copula-based dissimilarity measures. Specifically, a matrix of co-occurrences of the assets in the same cluster obtained from several partitions is derived. Such a matrix can lead to a more robust partition compared to the result from a single copula model or a specific hierarchical clustering linkage method. As a result, assets in the same cluster are expected to perform similarly during risky scenarios, and risk-averse investors could exploit this information to build a risk-diversified portfolio. An empirical demonstration of such a strategy is presented by using data from the EUROSTOXX50 index.

CP001 Room 43 POSTER SESSION Chair: Marios Demosthenous

C0327: **Applying regression methods to model survival data for gammarids (Amphipoda)**
Presenter: **Svitlana Shvydka**, Slovak University of Technology in Bratislava, Slovakia
Co-authors: Volodimir Sarabeev, Maria Zdimalova, Mykola Ovcharenko

The survival time of gammarids (Amphipoda) is considered using linear and nonlinear regression models with Gaussian distribution. The effect of individual parasite species richness (IPSR) on survival data under the influence of the host body length was analyzed. Three models were tested. For GLM and GAM models, the coefficients of the interaction terms were not significant. The marginal effects for both models revealed that the IPSR affected the survival time of gammarids measuring 7.7 to 15 mm in body length, but this influence was not observed in larger hosts. The smoother of Body length in GAM showed a non-linear effect. Gammarids around 11 mm have a higher probability of having a number of parasite species than smaller and larger hosts. The ANCOVA showed a significant difference in the effect of parasites on survival time between groups of hosts with body length of 7.7-15.1 and 15.1-21.8 mm. We evaluated the model fit by checking the level of deviance explained, analyzing the AIC and inspecting the residuals. S.S., V.S. are funded by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under project Nos.09I03-03-V01-00029, 09I03-03-V01-00017, respectively. The work of S.S. is partially supported by the grant VEGA1/0036/23.

C0458: **Monte Carlo simulated probabilistic forecasts of anonymized data based on probability-distributed ordinal classes**
Presenter: **Stefan Stroka**, ams Osram International GmbH, Germany

The growing interest in data privacy and anonymization poses challenges for unbiased and valid machine learning applications. Anonymization, such as through ordinal discretization, causes information loss due to coarsening. Current research suggests that a (pseudo-)continuous approach to ordinal classes can reduce this loss. The ordered classification of a metric target variable can be compared on a metric scale based on class boundaries. However, the accuracy of this approach depends on the number and size of the classes. A new method is introduced to reduce information loss and simulate probabilistic forecasts by modeling class distributions. Using a Gaussian mixture model (GMM) to model classes allows the replacement of discrete boundaries with probability distributions. Randomized samples from these distributions enhance the informational content for model training and enable prediction intervals through Monte Carlo simulations. This method is evaluated using an insurance dataset from "kaggle.com". The insurance amount, as the target variable, is divided into ordinal classes and used as training input along with other features. With a 10:90 training-test split, the methodology's performance is assessed with limited input data on a large test dataset, simulating a scenario where the exact insurance amount is known for only a small training sample.

C0491: **Geographically weighted logistic quantile regression**
Presenter: **Vivian Yi-Ju Chen**, National Chengchi University, Taiwan

Co-authors: Yu-Ting Lu

In spatial analysis, geographically weighted quantile regression (GWQR) has been proposed as a method for simultaneously capturing the heterogeneous relationships through the specification of varying regression parameters and modeling the response heterogeneity with the dependent variable modeled as a series of pre-specified quantiles. However, GWQR is designed only for unbounded continuous dependent variables. When handling bounded outcome data within a defined range, GWQR may yield biased results, such as incorrect inferences and out-of-range predictions. To address this limitation, an improved model is presented that integrates the concept of logistic quantile regression. The model specification of the proposed approach is presented, and relevant modeling issues are discussed, followed by an evaluation of its performance through simulations. Furthermore, the new approach is applied to a real dataset as an empirical illustration. The analysis results demonstrate that the proposed method provides more robust and informative insights compared to other existing analytical techniques.

Thursday 29.08.2024 10:30 - 12:30 Parallel Session I – COMPSTAT2024

CO121 Room 051 SMALL AREA ESTIMATION AND MIXED MODELS Chair: Domingo Morales

C0174: **Post-selection inference for fixed and mixed parameters in linear mixed models**
Presenter: **Stefan Sperlich**, University of Geneva, Switzerland
Co-authors: Katarzyna Reluga, Gerda Claeskens

While post-selection inference has received considerable attention in linear models, it is a neglected topic in the field of mixed models and mixed effect prediction. We developed methods and asymptotic theory for post-selection inference when the conditional Akaike information criterion was employed for model selection in a linear mixed model. These are used to construct confidence intervals for regression parameters, linear statistics and mixed effects under different scenarios, namely nested and general model sets as well as sets of misspecified models. The theoretical analysis is accompanied by simulation studies that confirm good performances. Moreover, they reveal a startling robustness of the classical confidence intervals for mixed parameters, which is in strong contrast to the findings for fixed parameters, indicating that random effects could automatically adjust for model selection. We illustrate our methodology along a study of the body mass index across different clusters in the US

C0236: **Random forests and mixed effects random forests for small area estimation of general parameters**
Presenter: **Nora Wuerz**, Otto-Friedrich-Universitaet Bamberg, Germany
Co-authors: Patrick Krennmair, Timo Schmid, Nikos Tzavidis

Random forests are highly effective for prediction due to minimal tuning parameters, automated model selection, and the ability to capture complex relationships. There is notable research on tree-based methods for survey data. More recently, theoretical properties of random forests have been explored for complex survey data. The focus is on random forests and extensions for estimating small area parameters, proposing mixed effects random forests to incorporate random effects crucial for small area estimation. This method extends prior work and uses a non-parametric bootstrap to correct the bias in the estimated residual variance before estimating the variance of the random effects. Estimators of general small area parameters are derived using area-specific smearing. For MSE estimation, a non-parametric block bootstrap with appropriate scaling of the residuals is used. Evaluation includes simulations and real data from poverty assessment in Mozambique, comparing forest-based estimators to industry standard methods like the Empirical Best Predictor. Findings highlight the impact of including random effects, the importance of data transformations, and the performance of estimators.

C0163: **Temporal M-quantile models and robust bias-corrected small area predictors**
Presenter: **Maria Bugallo**, Miguel Hernandez University of Elche, Spain
Co-authors: Domingo Morales, Nicola Salvati, Francesco Schirripa

In small area estimation, it is a smart strategy to rely on data measured over time. In this regard, linear mixed models are unable to properly capture time dependencies when the number of lags is large. Since there are no published studies addressing robust prediction in small areas based on time-dependent data, the M-quantile models are sought to be extended in this field of research. Indeed, the proposed methodology successfully addresses this challenge and offers flexibility to the widely imposed assumption of unit-level independence. Under the new model, robust bias-corrected predictors of small area linear indicators are derived. In addition, the optimal selection of the robustness parameter for bias correction is a theoretical contribution of the current research, exploring its applicability in outlier detection. As for the estimation of the mean squared error, a first-order approximation has been obtained under general conditions, and analytical estimators have been proposed. Several simulation experiments are carried out to investigate the performance of the new predictors and ensuing MSE estimators, as well as the optimal selection of the robustness parameters. Finally, an application to the Spanish Living Conditions Survey data is included to illustrate the usefulness of what has been done.

C0170: **Small area estimation of labour force indicators under bivariate Fay-Herriot model with correlated time and area effects**
Presenter: **Esteban Cabello**, Universidad Miguel Hernandez de Elche, Spain
Co-authors: Domingo Morales, Agustin Perez Martin

An area-level temporal bivariate linear mixed model is developed which incorporates correlated time effects for estimating socioeconomic indicators in small areas. The model is applied through the residual maximum likelihood method, deriving empirical best linear unbiased predictors for these indicators. Additionally, an approximation for the matrix of mean squared errors (MSE) is provided, and four MSE estimators are proposed. The first estimator involves a plug-in approach to the MSE approximation, while the remaining estimators are based on parametric bootstrap procedures. Three simulation experiments are conducted to assess the performance of the fitting algorithm, predictors, and MSE estimators. An application to real data from the 2016 to 2022 Spanish Living Conditions Survey is conducted. The focus is on estimating poverty proportions and gaps for the year 2022, categorized by provinces and sex.

C0481: **Improving small area poverty estimates with random-slope mixed models**
Presenter: **Naomi Diz-Rosales**, Universidade da Coruna, Spain
Co-authors: Maria Jose Lombardia, Domingo Morales

Nowadays, policymakers require detailed socio-economic indicators to assess poverty at very specific levels. However, the challenge arises when sample sizes are small, which affects the precision of estimates. Consequently, a small area estimation methodology is presented to derive predictors of the poverty proportion using random slope mixed models. A Poisson-type area model with random intercept and random slope is introduced, and bootstrap estimators of the mean squared error are defined, both with and without bias correction. A Laplace approximation algorithm is used to calculate maximum likelihood estimators of the model parameters and predictors of random effects. Through simulation experiments, the performance of the fitting algorithm, the predictors and the mean squared error estimators are evaluated. The optimal results obtained allow the model to be applied to real data from the Spanish Living Conditions Survey and the Spanish Labor Force Survey. The final objective is to estimate and map poverty proportions by province and sex in Spain, providing a valuable tool for decision-making in the allocation of resources and policies.

C0164: **Wildfire prediction using zero-inflated negative binomial mixed models: Application to Spain**
Presenter: **Domingo Morales**, University Miguel Hernandez of Elche, Spain
Co-authors: Maria-Dolores Esteban, Maria Bugallo

Zero-inflated negative binomial mixed models are adapted to wildfire data where the number of fires is zero in some months and is overdispersed in others because they allow to describe the patterns that explain both the number of fires and their non-occurrence, as well as provide good prediction tools. In addition to model-based predictions, a parametric bootstrap method is applied for estimating mean squared errors and constructing prediction intervals. This type of data is common in the Mediterranean, where the fires are mainly concentrated around the summer months. The statistical methodology and developed software are applied to model and predict the number of wildfires in Spain between 2002 and 2015 by provinces and months.

CO123 Room 052 DEPENDENCE MEASURES Chair: Marc-Oliver Pohle

C0396: **Universal copulas**
Presenter: **Gery Geenens**, University of New South Wales, Australia

Copulas are classically understood as cumulative distribution functions on the unit hypercube with standard uniform margins, referred to as "Sklar's copulas", owing to their central role in the decomposition of multivariate distributions established by the celebrated Sklar's theorem. The argument habitually put forward for outlining the appeal of copula models is that they allow pulling apart the dependence structure of a bivariate vector (the copula) from the individual behaviour of its marginal components. However, this interpretation can only be justified in the continuous framework, as copulas lose their "margin-free" nature outside of it, making Sklar's copula models unfit for modelling dependence between non-continuous variables. It is argued that the notion of copula should not be imprisoned in Sklar's theorem, and an alternative definition of copulas is proposed, which follows from approaching their role and meaning more broadly. This definition coincides with Sklar's copulas in the continuous framework but leads to different concepts in other settings. It is called construction universal copulas, and it is shown that these maintain all the pleasant properties (in particular, 'margin-freeness') that make Sklar's copulas sound and effective in continuous cases. The findings are illustrated with some examples of universal copula modelling between two discrete variables and between one continuous variable and one binary (Bernoulli) variable.

C0218: **Proper correlation coefficients for discrete random variables**
Presenter: **Jan-Lukas Wermuth**, Goethe University Frankfurt, Germany
Co-authors: Marc-Oliver Pohle

Contrary to Pearson correlation, Kendalls Tau and Spearmans Rho fulfill all important desirable properties for directed dependence measures, at least for continuous random variables. In the discrete case, however, they lose the property of attainability, meaning that they do not attain the values -1 and 1 under perfect negative and positive dependence and often severely understate the strength of dependence, impeding their usefulness as dependence measures. We show that their widely-used generalizations for the discrete case, Tau-B and Grade Correlation, can, to a certain extent, mitigate the severity of the attainability problem but are still non-attainable. We discuss attainable versions of rank correlations. For Spearmans Rho, we look into its several population definitions and demonstrate that an attainable version of it comes at the price of other shortcomings. For Kendalls Tau, on the contrary, we show that Goodman-Kruskals Gamma is a theoretically appealing and simple attainable generalization, which we consequently recommend as a suitable general-purpose dependence measure. We discuss further attainable versions of Tau and Rho and introduce an attainable version of Blomqvists Beta. In theoretical and empirical examples, we analyze and illustrate the attainability problem of classical correlation measures and how it is solved by attainable generalizations.

C0379: **A Regression Perspective on Generalized Distance Covariance and the Hilbert-Schmidt Independence Criterion**
Presenter: **Dominic Edelmann**, German Cancer Research Center, Heidelberg, Germany
Co-authors: Jelle Goeman

A seminal paper shows the equivalence of the Hilbert-Schmidt independence criterion (HSIC) and a generalization of distance covariance. The two notions of dependence are unified, with a third prominent concept for independence testing and the introduction of the global test. The new viewpoint provides novel insights into all three test traditions, as well as a unified overall view of the way all three tests contrast with classical association tests. As the main result, a regression perspective on HSIC and generalized distance covariance is obtained, allowing such tests to be used with nuisance covariates or for survival data. Several more examples of cross-fertilization of the three traditions are provided, involving theoretical results and novel methodology.

C0269: **Lancester correlation: A new dependence measure linked to maximum correlation**
Presenter: **Hajo Holzmann**, Philipps-Universitat Marburg, Germany
Co-authors: Bernhard Klar

Novel correlation coefficients are suggested that equal the maximum correlation for a class of bivariate Lancaster distributions while being only slightly smaller than the maximum correlation for a variety of further bivariate distributions. In contrast to maximum correlation, however, our correlation coefficients allow for rank - and moment-based estimators, which are simple to compute and have tractable asymptotic distributions. Confidence intervals resulting from these asymptotic approximations and the covariance bootstrap show good finite-sample coverage. In a simulation, the power of asymptotic and permutation tests for independence based on our correlation measures compares favorably to various competitors, including distance correlation and rank coefficients for functional dependence. Moreover, for the bivariate normal distribution, our correlation coefficients equal the absolute value of the Pearson correlation, an attractive feature for practitioners that is not shared by distance correlation, among others. We illustrate the practical usefulness of our methods in applications to two real data sets.

C0313: **Generalised covariances and correlations**
Presenter: **Marc-Oliver Pohle**, Heidelberg Institute for Theoretical Studies, Germany
Co-authors: Tobias Fissler

The covariance of two random variables measures the average joint deviations from their respective means. We generalise this well-known measure by replacing the means with other statistical functionals such as quantiles, expectiles, or thresholds. Deviations from these functionals are defined via generalised errors, often induced by identification or moment functions. As a normalised measure of dependence, a generalised correlation is constructed. Replacing the common Cauchy-Schwarz normalisation with a novel Fréchet-Hoeffding normalisation, we obtain the attainability of the entire interval $[-1,1]$ for any given marginals. We uncover favourable properties of these new dependence measures and establish consistent estimators. The families of quantile and threshold correlations give rise to function-valued distributional correlations, exhibiting the entire dependence structure. They lead to tail correlations, which should arguably supersede the coefficients of tail dependence. Finally, we construct summary covariances (correlations), which arise as (normalised) weighted averages of distributional covariances. We retrieve Pearson covariance and Spearman correlation as special cases. The applicability and usefulness of our new dependence measures are illustrated by demographic data from the Panel Study of Income Dynamics.

CO077 Room 43 FUSING MACHINE LEARNING AND STATISTICS Chair: Sonja Greven

C0247: **Adversarial random forests**
Presenter: **Marvin Wright**, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Germany

Adversarial random forests are presented, which are a provably consistent tree-based machine learning method designed for density estimation and generative modeling. Inspired by generative adversarial networks, our method employs a recursive procedure in which trees gradually learn structural properties of the data through alternating rounds of generation and discrimination. We achieve comparable or superior performance to state-of-the-art probabilistic circuits and deep learning models on various tabular data benchmarks while executing about two orders of magnitude faster on average. We provide an overview of the methodology, show benchmark results for density estimation and generative modeling, and introduce new methods for conditional sampling. The latter allows applications in missing data imputation and explainable AI, e.g., in conditional feature importance and counterfactual explanations.

C0211: **DoubleMLDeep: Estimation of causal effects with multimodal data**
Presenter: **Martin Spindler**, University of Hamburg, Germany

The use of unstructured, multimodal data, namely text and images, is explored in causal inference and treatment effect estimation. We propose a neural network architecture that is adapted to the double machine learning (DML) framework, specifically the partially linear model. An additional contribution is a new method to generate a semi-synthetic dataset, which can be used to evaluate the performance of causal effect estimation in the presence of text and images as confounders. The proposed methods and architectures are evaluated on the semi-synthetic dataset and compared to standard approaches, highlighting the potential benefit of using text and images directly in causal studies. Our findings have implications for researchers and practitioners in economics, marketing, finance, medicine and data science in general who are interested in estimating causal quantities using non-traditional data.

C0263: **Dropout regularization versus L2-penalization in the linear model**
Presenter: **Johannes Schmidt-Hieber**, University of Twente, Netherlands
Co-authors: Sophie Langer, Gabriel Clara

The focus is on the statistical behavior of gradient descent iterates with dropout in the linear regression model. In particular, non-asymptotic bounds for expectations and covariance matrices of the iterates are derived. In contrast with the widely cited connection between dropout and L2-regularization in expectation, the results indicate a much more subtle relationship, owing to interactions between the gradient descent dynamics and the additional randomness induced by dropout. We also study a simplified variant of dropout that does not have a regularizing effect and converges with the least squares estimator.

C0301: **Inference for semi-structured regression**
Presenter: **David Ruegamer**, LMU Munich, Germany
Co-authors: Rickmer Schulte, Thomas Nagler

In modern data analysis, neural networks offer significant flexibility, particularly for large and non-tabular data. Semi-structured models capitalize on this flexibility by defining a regression model with an additive predictor that combines a structured effect, such as a linear effect, with an unstructured effect represented by a neural network. While the most prominent approach currently involves training both parts jointly within a large, unified network, we discuss an alternative application of similar practical interest: Given a pre-trained deep neural network that was trained on another dataset, we explore under what assumptions we can derive inference statements for the structured model component when using the pre-trained model as the unstructured part in a semi-structured model.

C0302: **Deep nonparametric conditional independence tests for images**
Presenter: **Sonja Greven**, Humboldt University of Berlin, Germany
Co-authors: Marco Simnacher, Xiangnan Xu, Hani Park, Christoph Lippert

Conditional independence tests (CITs) are often used to study the relationship between health outcomes and exposures or genetic information, adjusting for potential confounders. While complex health outcomes are increasingly studied using medical imaging, there is a significant gap in the literature regarding CITs for an image and a scalar, given a confounding vector. To fill this gap, we introduce novel deep nonparametric CITs (DNCITs), which integrate nonparametric CITs for vector-valued data with embedding maps to extract feature representations from images. We show the validity of DNCITs under conditions on the embedding map that are fulfilled, particularly for learning schemes such as sample splitting, transfer learning, and (un)conditional unsupervised learning. We discuss the transferability of existing nonparametric CITs to our situation and propose tests adapted to our setting. We provide an implementation for these DNCITs in an R package and study their validity, power and sensitivity to different factors in an extensive simulation study. We apply the DNCITs to study the dependence between brain MRI scans and behavioral traits, given confounders, in healthy individuals in the UK Biobank, to shed new light on previous mixed results from several personality neuroscience studies with our more powerful tests on a larger dataset.

CO105 Room 44 TEXT MINING: METHODS AND APPLICATIONS Chair: Anna Staszewska-Bystrova

C0153: **Analysing the impact of removing infrequent terms on topic quality in LDA models**
Presenter: **Viktoriia Naboka-Krell**, Justus Liebig Unversity of Giessen, Germany
Co-authors: Peter Winker, Victor Bystrov, Anna Staszewska-Bystrova

An initial procedure in text-as-data applications is text preprocessing. One of the typical steps, which can substantially facilitate computations, consists of removing infrequent words believed to provide limited information about the corpus. Despite the popularity of vocabulary pruning, not many guidelines on how to implement it are available in the literature. The aim is to fill this gap by examining the effects of removing infrequent words for the quality of topics estimated using Latent Dirichlet Allocation. The analysis is based on Monte Carlo experiments taking into account different criteria for infrequent terms removal and various evaluation metrics. The results indicate that pruning is beneficial and that the share of vocabulary which might be eliminated can be quite considerable.

C0204: **Textual content and academic journals selectivity: A case of economic journals**
Presenter: **Pawel Baranowski**, Institute of Economic and Financial Research, Lodz, Poland, Poland
Co-authors: Szymon Wojcik

A large supply of papers obstructs the editorial procedures in scientific journals, especially in top-quality academic journals. Moreover, this phenomenon 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 can help distinguish the papers worth reading by the editor from those whose scientific quality does not meet the standards. 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 500 academic papers representing economics and finance or business. The papers were collected from journals of three levels of selectivity, namely: highly selective (top-tier journals), moderately selective (journals listed on the DOAJ list), and non-selective (predatory journals). More specifically, we applied both pre-trained and fine-tuned Sci-BERT model on anonymised texts of academic papers. The results show that the pure textual content may give over 80% out-of-sample accuracy in classifying texts into the three levels of selectivity. The outcomes prove the usefulness of NLP in distinguishing the scientific quality of the paper and support Bealls classification of predatory journals.

C0213: **Automated question answering for unveiling leadership dynamics in U.S. presidential speeches**
Presenter: **Krzysztof Rybinski**, Vistula University Warsaw, Poland

An innovative methodology is introduced utilising deep learning and automated question-answering techniques to explore leadership dynamics. The research analyses 989 speeches from all U.S. Presidents, applying a machine learning model to decipher the essence of effective leadership within the context of presidential rhetoric. This approach facilitates an interrogation of "What constitutes a great leader?" by extracting pertinent attributes from these presidential discourses. The study conducts a comprehensive statistical examination of the responses generated across historical epochs. Furthermore, the research employs a regression analysis that extends over 120 years, integrating the attributes of outstanding leadership with the Economic Policy Uncertainty (EPU) index specific to the United States. The findings indicate that periods marked by heightened uncertainty necessitate leaders with a charismatic approach, whereas servant leadership is more effective during times of reduced uncertainty. After an extensive validation and robustness assessment, it was concluded that the outcomes are steadfast, notwithstanding variations in key parameters of the deployed machine learning models. Moreover, these findings align coherently with the qualitative assessments of U.S. Presidential views on leadership undertaken in prior research, thus contributing novel insights into the intricate relationship between leadership qualities and historical and economic contexts.

C0216: **Goodness-of-fit testing in topic models**
Presenter: **Anna Staszewska-Bystrova**, University of Lodz, Poland

Co-authors: Victor Bystrov

Topic models used for structural analysis of textual data are most often evaluated on the basis of characteristics of extracted topics. Apart from providing coherent topics, these models should also exhibit a good fit to the data. The standard goodness-of-fit tests are not suited for large corpora that are characterized by a sparse distribution of terms. We propose a testing procedure that relies on averaging of goodness-of-fit statistics across documents in a corpus. The performance of the tests is evaluated in the latent Dirichlet allocation (LDA) model by means of Monte Carlo simulations under the assumption of known parameters. A bootstrap procedure for goodness-of-fit testing in the estimated LDA is also proposed, and the size and power of the bootstrap tests are analysed.

C0370: **Startups in African agriculture sector: Insights from the computational linguistics**

Presenter: **Nouhaila Belaid**, Africa Business School (ABS)-Mohammed VI Polytechnic University (UM6P), Morocco

Co-authors: Ivan Savin

Africa's agriculture sector holds promise for sustainable development, which is of utmost priority given the current challenges in population growth, resource scarcity and climate variability. Startups, especially in agriculture technology, are key drivers of innovation. Employing structural topic modeling, textual descriptions of 36,232 African startups in the Crunchbase database are studied. A new classification of startups in Africa is offered that includes 30 topics ranging from real estate and telecom through healthcare and education to cybersecurity and cryptocurrency. While the Crunchbase database has an internal classification, companies are attributed to classes by individual company owners, resulting in inconsistencies, gaps and inaccuracies. The approach allows the mitigation of these challenges. Startups are found in this continent, demonstrating a trend towards more cross-industry operations, i.e. recombining more business topics in one company. For example, many companies work on the intersection of agriculture and blockchain, farming and insurance, or healthcare and AI. It is also found that topics like real estate, manufacturing and AI are concentrated in South and Northern Africa, while Eastern Africa attracts more startups in renewable energy and agriculture and West Africa for online marketplaces and e-commerce. Topics that attract the most venture capital investments are mining, telecom, renewable energy and startup incubators.

CO102 Room 45 CLUSTERING AND CLASSIFICATION FOR COMPLEX DATA Chair: Mika Sato-Ilic

C0265: **The use of mixture models for clustering data with structured dependence**

Presenter: **Shu-Kay Angus Ng**, Griffith University, Australia

Co-authors: Richard Tawiah, Geoffrey McLachlan

Identifying (disadvantaged) subgroups is fundamental and decisive in solving many real-world problems. Mixture models underpin a variety of statistical methods in cluster and latent class analyses for finding subgroups, outliers, and distinctive features between subgroups. Statistical inference for mixture models assumes that observed data are independent of one another. However, modern study designs often generate data structures with non-negligible dependence among data (e.g., patients treated in a hospital share the same hospital effect in multilevel studies). Thus, the independence assumption becomes invalid. Clustering methods (or mixture models) that ignore the structured dependence (by assuming zero hospital effect) can overlook the significance of such effect and data variability, resulting in misleading findings or failure to identify important risk factors. We present a statistical framework in mixtures of generalised linear mixed models (GLMMs) for clustering data with complex structured dependence. We introduce random-effect modelling techniques that can effectively capture complex intra- and between-subject correlations among observations due to various forms of dependence. An efficient estimation of model parameters is achieved using extended best linear unbiased prediction (BLUP) and approximate residual maximum likelihood (REML) procedures. We consider several data sets to show the capacity of this clustering approach.

C0300: **On Lasso Poisson regression for categorical variables**

Presenter: **Mariko Yamamura**, Hiroshima University, Japan

Co-authors: Mineaki Ohishi, Hirokazu Yanagihara

One of the main advantages of the Lasso is that it provides estimation results with zero coefficients. This means that explanatory variables with estimated coefficients of zero are not included in the model. The case to be considered is that of a categorical explanatory variable. When estimation using Lasso is performed on a categorical variable, some of the multiple categories in the categorical variable will be estimated as zero. This does not mean that the categories estimated as zero are not included in the model but rather that they are chosen as a baseline for the categorical variables. Since the estimates of the coefficients for each category of a categorical variable and the interpretation of these estimates depend on the baseline, it is crucial to identify the baseline category. The aim is to identify which categories are estimated to be zero. The objective function is a Lasso-Poisson regression in which all the explanatory variables are categorical. Theoretical clarifications and numerical experiments show that the baseline corresponds to a category with a coefficient equal to the weighted median of the coefficients of the categorical variable.

C0330: **Clustering based multidimensional scaling for mixed data**

Presenter: **Mika Sato-Ilic**, University of Tsukuba, Japan

Multidimensional scaling (MDS) is a well-known method for dimensional reduction and visualization of multidimensional data. Clustering-based MDS is proposed for data of the mixed types, which comprises numerical and categorical data regarding quantitative and qualitative variables. For this type of data, the main difficulty for the treatment of data depends on the difference in quantity and quality of intrinsic information contained in the data. The amount of data information is larger for the numerical data; therefore, traditionally, the transformation from the numerical data to categorical data by aggregating the data information for the categorical data has been used. However, in this case, complex methods are needed for the aggregation of the transformation. Therefore, a simple technique which can treat both data types through the same data projection into the lower dimensional space is proposed. This method includes the classification structure obtained by using fuzzy clustering; the categorical data part of the mixed data can be summarized and expresses the simple relationship with the objects of the numerically obtained data part of the mixed data.

C0338: **Applicability of TreeSHAP to analyze real estate data**

Presenter: **Koki Kirishima**, Hiroshima University, Japan

Co-authors: Mineaki Ohishi, Ryoya Oda, Kensuke Okamura, Yoshimichi Itoh, Hirokazu Yanagihara

Machine learning models such as random forests are often used to analyze real estate data. Although machine learning models achieve very high predictive accuracy, the difficulty lines interpreting the predictions. The SHAP, which is one of the methods of XAI, has been proposed to overcome this difficulty and facilitates the interpretation of the predictions by linearly decomposing the influence of each explanatory variable on the predictions. For random forests, TreeSHAP, which allows SHAP to be computed strictly and quickly, has been proposed. The applicability of TreeSHAP to analyze real estate data is discussed.

C0356: **Classification method for corrupted label data using density ratio**

Presenter: **Masaaki Okabe**, Doshisha University, Japan

Co-authors: Hiroshi Yadohisa

During the training of the classification model, the labels of the training data were assumed to be correct. However, human errors such as mislabeling can result in incorrect labels being assigned to objects. If training data contain incorrect labels, the classification accuracy of the model may be reduced. A prior study proposed a classification method that uses the balanced error rate as the objective function for training from corrupted label data. This method assumes independence between features and corrupted occurrences, given a true label, and specifically addresses cases where mislabeling occurs randomly. However, classification may not work well when label corruptness is correlated with features. If label errors depend on the features, they are correlated with the features. The focus is on the arrangement of Menon's assumptions, proposing an estimation of classification models with relaxed assumptions compared to existing methods by using the ratio of the contaminated label distribution to the true label distribution of the obtained data.

CC132 Room 001 FINANCIAL RISK Chair: Alessandra Amendola

C0289: **Extracting macro factors in bond risk premia using a supervised method**
Presenter: **Hiroyuki Kawakatsu**, Dublin City University, Ireland

The aim is to re-examine whether yield factors for bond risk premia span macroeconomic variables. Rather than commonly used macroeconomic variables, such as output growth and inflation, the space spanned by a large number of macroeconomic variables is considered. A small number of factors extracted from a supervised method is shown to be priced and not spanned by other well-known and commonly used factors in the literature.

C0499: **Semi parametric financial risk forecasting incorporating multiple realized measures**
Presenter: **Rangika Peiris**, University of Sydney Business School, Australia
Co-authors: Chao Wang, Richard Gerlach, Minh-Ngoc Tran

A semi-parametric joint value at risk (VaR) and expected shortfall (ES) forecasting framework employing multiple realized measures are developed. The proposed framework extends the realized exponential GARCH model to be semi-parametrically estimated via a joint loss function whilst extending quantile time series models to incorporate multiple realized measures. A quasi-likelihood is built, employing the asymmetric Laplace distribution directly linked to a joint loss function, enabling Bayesian inference for the proposed model. An adaptive Markov chain Monte Carlo method is used for the model estimation. The empirical section evaluates the performance of the proposed framework with six stock markets from January 2000 to June 2022, covering the period of COVID-19. Three realized measures, including 5-minute realized variance, bi-power variation, and realized kernel, are incorporated and evaluated in the proposed framework. One-step-ahead VaR and ES forecasting results of the proposed model are compared to a range of parametric and semi-parametric models, lending support to the effectiveness of the proposed framework.

C0276: **Diversifying risk parity portfolios with high-frequency principal components**
Presenter: **Laura Garcia-Jorcano**, Universidad de Castilla-La Mancha, Spain
Co-authors: Massimiliano Caporin, Juan-Angel Jimenez-Martin

The diversified risk parity (DRP) strategy is used for multi-asset allocation to generate diversified equity portfolios. Creating uncorrelated risk sources by means of high-frequency principal components analysis (HF-PCA), we obtain maximum diversification portfolios when equally budgeting risk to each of the uncorrelated risk sources. We forecast the risk factors and trace the role of firms/industries as potential sources of financial risk in different periods of time. The empirical analysis carried out using one-minute returns of stocks included in the S&P 100 index from 2003 to 2022 belonging to ten industry groups, shows that compared to classical risk-based allocation schemes, the DRP strategy provides the most convincing risk-adjusted performance and the most diversified portfolio among the investigated alternatives according to several concentration indices and risk decomposition characteristics. HF-PCA allows the DRP strategy to constantly adapt to changes in risk structure and maintain a balanced exposure to the prevailing uncorrelated risk sources. This tool can help a portfolio manager understand and choose risk sources that have earned risk by focusing on those risk factors.

C0390: **Evaluating financial tail risk forecasts with the model confidence set**
Presenter: **Lukas Bauer**, University of Freiburg, Statistics and Econometrics, Germany

The focus is on the first provision of results on the finite sample properties of the model confidence set (MCS) applied to the asymmetric loss functions specific to financial tail risk forecasts, such as value-at-risk (VaR) and expected shortfall (ES). The emphasis is on statistical loss functions that are strictly consistent. The comprehensive simulation results show that, first, the MCS test keeps the best model more frequently than the confidence level $1-\alpha$ in most settings. Second, it eliminates a few inferior models for out-of-sample sizes of up to four years. Third, the MCS test shows little power against models that underestimate tail risk at the extreme quantile levels p=0.01 and p=0.025, while the power increases with the quantile level p. These findings imply that the MCS test may be suitable to narrow down a set of competing models but that it is not appropriate to test if a new model beats its competitors due to the lack of power.

C0475: **US equity announcement risk premia**
Presenter: **Lukas Petrasek**, Charles University Prague, Czech Republic
Co-authors: Jiri Kukacka

The announcement risk premia is analyzed on the US market. Previous studies have found that a significant portion of the overall risk premia is earned on FOMC meeting days and when inflation and employment reports are published. The evidence suggests that while the announcement risk premium for these days still exists, there is a much wider range of macroeconomic data releases to consider. It is found 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 are 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. It shows how premia changes across different industries. A trading strategy that holds long positions in equities on announcement days and long positions in risk-free assets on non-announcement days has a more than two times higher Sharpe Ratio over a simple buy-and-hold strategy on equities. It also documented how the risk premia of well-established asset pricing factors, e.g., beta and size, changes between announcement and non-announcement days. These results are robust to the inclusion of several controls and are both economically and statistically significant.

CC022 Room 050 VARIABLE AND MODEL SELECTION Chair: Florian Frommlet

C0234: **AIC for many-regressor heteroskedastic regressions**
Presenter: **Stanislav Anatolyev**, CERGE-EI and New Economic School, Czech Republic

The original and corrected Akaike information criteria (AIC) have been routinely used for model selection for ages. The penalty terms in these criteria are tied to the classical normal linear regression, characterized by conditional homoskedasticity and a small number of regressors relative to the sample size. We derive, from the same principles, a general version that takes account of conditional heteroskedasticity and regressor numerosity. The new AICm penalty takes the form of a ratio of certain weighted average error variances and can be operationalized via unbiased estimation of individual variances. The feasible AICm criterion still minimizes the expected Kullback-Leibler divergence up to an asymptotically negligible term that does not relate to regressor numerosity. In simulations, the feasible AICm does select models that deliver systematically better out-of-sample predictions than the classical criteria.

C0473: **Marginalized LASSO in the difference-based partially linear model for variable selection**
Presenter: **Mina Norouzirad**, Center for Mathematics and Applications (NovaMath), Portugal
Co-authors: Ricardo Moura, Mohammad Arashi, Filipe Marques

The difference-based partially linear model is suitable for regression when both linear and nonlinear predictors are present in the data. Optimizing the weights using the difference-based method presents challenges in variable selection, particularly with low-variance predictors. A novel methodology based on marginal theory is proposed to address these mixed relationships effectively, emphasizing variable selection through a marginalized LASSO estimator with a penalty term that is less severe and related to the difference order. Comprehensive simulation experiments evaluate the performance of the proposed technique in estimation and prediction compared to the LASSO estimator. Additionally, the bootstrapped method is employed to assess the performance of the proposed prediction method using the King House dataset.

C0463: **Variable selection and estimation in non-linear mixed-effects models in high dimensional setting**

Presenter: **Antoine Caillebotte**, INRAE, France

Co-authors: Estelle Kuhn, Sarah Sarah Lemler

Mixed-effects models are a robust and increasingly popular tool for statistical modeling, in particular for the analysis of repeated measurements and longitudinal data. A non-linear mixed-effects model is considered, including high-dimensional covariates at the individual parameter modeling level. The focus is on variable selection and parameter estimation in this model. To handle the high dimensional setting and select a subset of relevant covariates, a LASSO type penalized maximum likelihood estimate is considered. The expectation maximization (EM) algorithm is a classical method for performing inference in mixed-effects models. However, it has several practical and theoretical limitations, such as the usual assumption that the model belongs to the exponential family. To circumvent this assumption, the use of the efficient Fisher-preconditioned stochastic gradient descent (Fisher-SGD) algorithm is proposed, which enables maximum likelihood inference in very general latent variable models. A proximal operator is added to this algorithm, which allows for penalized likelihood maximization and effective variable selection in high-dimensional contexts. The proposed algorithm's performance is illustrated in inference and variable selection through numerical simulations. The methodology is applied to a biological real dataset of wheat leaf senescence.

C0262: **Comparison of the LASSO and IPF-LASSO methods for multi-omics data: Variable selection with Type I error control**

Presenter: **Charlotte Castel**, Oslo University Hospital, Norway

Variable selection in high-dimensional regression modelling involving omics data is a hard problem, and establishing robust and dependable methods is essential. The IPF-LASSO model has advanced this field by allowing integration of diverse omics modalities, introducing distinct penalty parameters for each modality. However, controlling false positives when incorporating these heterogeneous data layers remains an unresolved challenge. To address this problem, we used stability selection for variable selection with error control. We applied stability selection to both the LASSO and IPF-LASSO, and the objective was to evaluate if the modality-specific penalties in the IPF-LASSO increase statistical power while maintaining error control. Analyses were conducted on two high-dimensional datasets, characterized by independent and correlated variables, respectively. Simulation studies indicated that while both methods were able to control false positives, IPF-LASSO increased power, especially under conditions with distinct differences in the relevance of variables across modalities. The different models were illustrated using data from a study on breast cancer treatment, where the IPF- LASSO model was able to select some highly relevant clinical variables. To our knowledge, this is the first study to integrate multiple correlated omics data modalities into a regression framework while controlling false positives.

C0387: **Shapley values for regression models with interactions**

Presenter: **Mark van de Wiel**, Amsterdam University Medical Centers, Netherlands

In machine learning, Shapley values are popular variable importance metrics, as they provide a unique combination of properties, such as efficiency and linearity, which enhances their use and interpretability. For regression models with interactions, quantification of variable importance is not trivial either, as the variable's contribution is present in multiple terms. The use of Shapley values in this context is argued, and a computationally efficient formula is derived to compute those for the model. Importantly, when applying appropriate shrinkage, it is shown that appropriate inference is available via credible intervals. The Shapley values are illustrated in a large epidemiological study. First, the regression model is demonstrated with two-way interactions outperforming a regression model with main effects only and a random forest in terms of prediction in a setting with sample size $n = 1000$, $p = 14$ variables and $q = 85$ two-way interactions. Hence, the model is a good candidate for further use. Then, the Shapley values and their uncertainties illustrate how variable importance differs across individuals due to the interaction terms. It visualizes how the Shapley value nicely decomposes into contributions of the main effect and the interactions, which allows the assessment of the relative importance of these effects. All in all, the aim is to show that Shapley values are a useful addendum to the statistician's toolbox for interpreting non-trivial regression models.

Thursday 29.08.2024 14:00 - 15:30 Parallel Session J – COMPSTAT2024

CI008 Room 44 MEASURE TRANSPORTATION AS A TOOL FOR STATISTICAL INFERENCE Chair: Gery Geenens

C0215: **Distribution-free tests of multivariate independence based on center-outward signs and ranks**
Presenter: **Hongjian Shi**, Technical University of Munich, Germany
Co-authors: Marc Hallin, Mathias Drton, Fang Han

Rank correlations have found many innovative applications in the last decade. In particular, suitable rank correlations have been used for distribution-free and consistent tests of independence between random variables. However, it has long remained unclear how one may construct distribution-free yet consistent tests of independence between random vectors since the traditional concept of ranks relies on ordering data and is tied to univariate observations. We will discuss how this problem can be addressed via a general framework for designing multivariate dependence measures and associated test statistics based on the recently developed notion of center-outward ranks and signs, a multivariate generalization of traditional ranks. We obtain multivariate extensions of Hajek asymptotic representation and use them to conduct local power analyses that demonstrate the statistical efficiency of our tests. I will also present multivariate extensions of the quadrant, Spearman, Kendall, and van der Waerden tests based on center-outward ranks and signs. A multivariate Chernoff-Savage property is provided to guarantee that, under elliptical generalized Konijn models, the asymptotic relative efficiency of our van der Waerden tests with respect to Wilks' classical (pseudo-)Gaussian procedure is strictly larger than or equal to one.

C0306: **Nonparametric multiple-output center-outward quantile regression**
Presenter: **Alberto Gonzalez Sanz**, Toulouse III, France
Co-authors: Marc Hallin, Eustasio del Barrio

Based on the novel concept of multivariate center-outward quantiles, the problem of nonparametric multiple-output quantile regression is considered. The approach defines nested conditional center-outward quantile regression contours and regions with given conditional probability content irrespective of the underlying distribution; their graphs constitute nested center-outward quantile regression tubes. Empirical counterparts of these concepts are constructed, yielding interpretable empirical regions and contours that are shown to consistently reconstruct their population versions in the Pompeiu-Hausdorff topology. Our method is entirely non-parametric and performs well in simulations including heteroskedasticity and nonlinear trends; its power as a data-analytic tool is illustrated on some real datasets.

C0274: **Quantiles and quantile regression on Riemannian manifolds: A measure-transportation-based approach**
Presenter: **Hang Liu**, University of Science and Technology of China, China
Co-authors: Marc Hallin, Thomas Verdebout

Increased attention has been given recently to the statistical analysis of variables with values on manifolds. A natural but nontrivial problem in that context is: "can we define quantile concepts for such variables?" We are proposing a solution to that problem for compact Riemannian manifolds without boundaries; typical examples are polyspheres, hyperspheres, and toroidal manifolds equipped with Riemannian metric. Our concept of quantile functions comes along with a concept of distribution function and, in the empirical case, ranks and signs. The absence of a canonical ordering is offset by resorting to the data-driven ordering induced by optimal transports. Statistical inference applications, from GOF to distribution-free rank-based testing, are without number. Of particular importance is the case of quantile regression with directional or toroidal multiple output, which is given special attention. Theoretical properties, such as the uniform convergence of the empirical distribution and conditional (and unconditional) quantile functions and distribution-freeness of ranks and signs, are established. Extensive simulations are carried out to illustrate these novel concepts.

CO104 Room 051 HIGH DIMENSIONAL DATA ANALYSIS FOR SOCIAL SCIENCES Chair: Ida Camminatiello

C0282: **Inside the black-box models through explainable decision tree ensembles**
Presenter: **Carmela Iorio**, University of Naples, Federico II, Italy
Co-authors: Agostino Gnasso, Massimo Aria

Models that can predict outcomes and explain the process by which they are produced are urgently needed in the social sciences. Explainable machine learning is about making the inner workings of models easy to understand, from input to output. Ensemble methods are popular because they combine multiple models to achieve accurate solutions. Thanks to its impressive ability to accurately predict outcomes, Random Forest (RF) is a common tool for regression and classification problems. Despite their apparent simplicity, RF models are often perceived as black-box models due to the complexity of the decision trees they generate. We have developed a solution to this problem: Explainable Ensemble Trees. This methodology provides explainable decision trees within the RF framework. It offers both predictive performance and a visual representation that is intuitively understandable. The aim is to represent the relationships between variables to improve explainability. These models are intended to explain decision processes occurring in domains where results may have important consequences. Acknowledgment: This research has been financed by the following research projects: PRIN-2022 "SCIK-HEALTH" (Project Code: 2022825Y5E; CUP: E53D2300611006); PRIN-2022 PNRR "The value of scientific production for patient care in Academic Health Science Centres" (Project Code: P2022RF38Y; CUP: E53D23016650001)

C0310: **Composite indicators to deep diving into residents perceptions of tourism**
Presenter: **Pasquale Sarnacchiaro**, University of Naples Federico II, Italy
Co-authors: Irene Ariante

Tourism represents a multi-faced phenomenon encompassing economic, social-cultural, and environmental dimensions. Residents ' perceived impacts play a crucial role in promoting sustainable tourism, and the measure of these perceptions represents strategic information for policymakers. A survey was conducted to study residents' perceptions. In particular, a questionnaire composed of 27 questions related to the three dimensions of tourism and the demographic characteristics of respondents was administered to a sample of 337 residents from the historical centre of Naples. The collected data were used for the construction of four composite indicators: an overall tourism perception indicator and three specific composite indicators for each dimension. Higher-order factor analysis has been used to estimate the indicators. Furthermore, an integrated approach of the multi-group analysis and the analysis of means (ANOM) allowed us to explore differences in perceptions between distinct sub-groups within the residents. This approach provides an understanding of significant variations in the perceptions of different clusters by offering valuable insights into the factors influencing different views on tourism.

C0359: **A study for reducing the economic inequalities among European countries**
Presenter: **Alfonso Piscitelli**, Federico II University of Naples, Italy
Co-authors: Ida Camminatiello

The aim is to address the issue of economic inequalities among European countries, which is in line with the broader agenda of sustainable development outlined by the United Nations. A regression model will be carried out to analyse the socio-economic determinants influencing economic inequalities. The socio-economic literature suggests that the inequality measure has memory, i.e., the state at the current time depends on the state at earlier times. Several time series models have been proposed to deal with processes with memory. Such models are basically formulated as ordinary least squares regression but include lagged dependent and independent variables. Given the high-dimensional nature of the data and the limited sample size, a partial least squares regression is considered as the most suitable statistical methodology. Incorporating time series analysis into the same framework of modelling non-dynamic data can contribute to reducing complexity and difficulties, facilitating the applicability of the model for informing policy decisions and promoting sustainable development.

C0316: **Italian subjective well-being: A territorial and longitudinal analysis through a poset methodology**
Presenter: **Enrico Ivaldi**, Iulm University, Italy
Co-authors: Leonardo Salvatore Alaimo

The study and measurement of subjective well-being has a long tradition in the literature. The various research studies have focused both on defining the concept of subjective well-being and on constructing a measure of it. In recent decades, research interest has grown in the synthesis of indicators relating to subjective well-being with the aim of constructing a synthetic measure of subjective well-being. By using data about citizens collected through the Multi-purpose Surveys on families conducted by Istat (Institute Italian National Statistics Institute), the purpose is to define a synthetic index, which allows the comparison of different individuals and different Italian Regions about their level of subjective well-being. The study analyses the longitudinal data produced by Istat between 2014 and 2021 with the aim of understanding if and how subjective well-being has changed in recent years. In summarizing the information, the Partially Ordered Sets methodology is applied, which allows respect for the multidimensionality of the phenomenon and the ordinal nature of the indicators used.

CO081 Room 052 FINANCIAL ECONOMETRICS Chair: Roxana Halbleib

C0196: **Local predictability in high dimensions**
Presenter: **Philipp Adaemmer**, University of Greifswald, Germany
Co-authors: Sven Lehmann, Rainer Alexander Schuessler

A novel time series forecasting method is proposed, which is designed to handle vast sets of predictive signals, many of which are irrelevant or short-lived. The method transforms heterogeneous scalar-valued signals into candidate density forecasts via time-varying coefficient models and, subsequently, combines them into a final density forecast via time-varying subset combination. The approach is computationally fast, because it uses online prediction and updating. We validate our method through simulation analyses and apply it to forecast daily aggregate stock returns as well as quarterly inflation, using over 12,000 and over 400 signals, respectively. We find superior forecasting performance and lower computation time for our approach compared to competitive benchmark methods.

C0228: **Predicting Value at Risk for cryptocurrencies with generalized random forests**
Presenter: **Rebekka Buse**, Karlsruhe Institute of Technology, Germany
Co-authors: Melanie Schienle

The prediction of Value at Risk (VaR) for cryptocurrencies is studied. In contrast to classic assets, returns of cryptocurrencies are often highly volatile and characterized by large fluctuations around single events. Analyzing a comprehensive set of 105 major cryptocurrencies, we show that Generalized Random Forests (GRF) adapted to quantile prediction have superior performance over other established methods such as quantile regression, GARCH-type and CAViaR models. This advantage is especially pronounced in unstable times and for classes of highly-volatile cryptocurrencies. Furthermore, we identify important predictors during such times and show their influence on forecasting over time. Moreover, a comprehensive simulation study also indicates that the GRF methodology is at least on par with existing methods in VaR predictions for standard types of financial returns and clearly superior in the cryptocurrency setup.

C0237: **Testing for a breakdown in forecast accuracy under long memory**
Presenter: **Philipp Sibbertsen**, University of Hannover, Germany
Co-authors: Jannik Kreye

A test is proposed to detect a forecast accuracy breakdown in a long memory time series. The proposed method uses a double sup-Wald test against the alternative of a structural break in the mean of the out-of-sample squared error loss series. To address the problem of estimating the long-run variance under long memory, a robust estimator is applied. The breakpoint is determined by a long memory robust CUSUM test. We provide theoretical and simulation evidence on the memory transfer from the time series to the forecast residuals. The finite sample size and power properties of the test are derived in a Monte Carlo simulation. We find that only the fixed forecasting scheme leads to a monotonic power function. The practical relevance of the method is demonstrated by detecting a forecast failure in German electricity prices.

C0278: **High dimensional change point estimation onthe community structure of networks**
Presenter: **Yarema Okhrin**, Universitaet Augsburg, Germany

Community detection is one of the cornerstones or building blocks in statistics and, more broadly, in data science research. The well-known models include K-means, stochastic block model, LDA, and spectral clustering model, and many others. All detection methods strive to a classification problem and assume a steady community structure. As networks evolve, the structure of the network, including degree, centrality, and the underlying communities, may change. We develop an offline monitoring technique aimed at detecting changes in community assignments. The approach is based on the ratio of eigenvectors and CUSUM aggregation. We prove the consistency of the estimated change point and extend it to a multiple change-point setting. We run an extensive simulation study and compare the results to the well-established benchmarks. The empirical study confirms the efficiency of the algorithm.

CO079 Room 43 NLP APPLICATIONS IN SOCIAL SCIENCES Chair: Ivan Savin

C0182: **Identification of innovation drivers based on technology-related news articles**
Presenter: **Albina Latifi**, Justus Liebig University Giessen, Germany
Co-authors: Peter Winker, David Lenz

Innovations contribute to economic growth. Hence, knowledge about drivers of innovation activities is a necessary input for economic policymaking when it comes to implementing targeted support measures. We focus on firms as potential drivers of innovation and use a novel data-driven approach to identify them. The approach is based on news articles from a technology-related newspaper for the period 1996-2021. In the first step, natural language processing (NLP) tools are used to identify latent topics in the text corpus. Expert knowledge is used to tag innovation-related topics. In the second step, a named entity recognition (NER) method is used to detect firm names in the news articles. Combining the information about innovation-related topics and firms mentioned in news articles linked to these topics provides a set of firms linked to each innovation-related topic. The results suggest that the approach helps identify drivers of innovation activities going beyond the usual suspects. However, given that the rate of false alarms is not negligible, in the end also, human judgement is needed when using this approach.

C0335: **The diffusion of green energy narratives**
Presenter: **Jessica Birkholz**, University of Bremen, Germany
Co-authors: Philip Kerner

What are the important factors for regional path dependence in the energy transition? The purpose is to address this question by considering the spread of shared narratives that guide expectations as a specific potential factor of regional development. Recent theories of decision-making highlight the pivotal role of narratives for actors to make investment decisions in the face of uncertain futures. Based on this theory, the focus is on how regional narratives and related uncertainties on renewable energy spread across space and time in German regions. First, to identify shared narratives, press release data is used, which provides a wide breadth of topics and a high granularity in the within-and between-region dimensions.

These data are analyzed with natural language processing techniques to characterize narrative content and sentiment. Second, to assess the diffusion of these narrative measures, a spatiotemporal diffusion model is estimated that facilitates jointly estimated spatial lags, temporal lags, and common (nationwide) factors. The contribution to the literature is in several ways. First, it explores a recently discovered data source further by concentrating on specific topics relevant to renewable energy. Second, it provides an understanding of the diffusion patterns and processes of renewable energy narratives in German regions.

C0352: **Effectiveness of mandatory CSR reporting: Improvement of firms' CSR disclosure information**
Presenter: **Bianca Minuth**, ESCP Business School, Paris, France

In a world of climate change and increasing social incidence more and more jurisdictions are changing their regulations towards a mandatory corporate social responsibility (CSR) reporting regime. Likewise, the CSR Directive 2014/95/EU was issued in 2014 with the objective to increase transparency of non-financial information and thus improving CSR disclosure quality in the European Union (EU). I examine whether firms adjust their CSR disclosure content as a response to the CSR disclosure mandate. Applying topic modeling, I investigate firms' disclosure quality, comparability, and reliability around the CSR Directive. In particular, I make use of the BERTopic modeling technique as a state-of-the-art large language model (LLM) to cluster CSR-related topics and analyse their development and similarity over time. Using a difference-in-difference model, I compare European firms CSR disclosure content with a sample of U.S. control firms before and after the introduction of a CSR regulation. The results provide evidence of an improvement in firms CSR disclosure quality and comparability after the CSR Directive. With my study, I enrich the debate on CSR reporting regime choices towards a more sustainability-oriented society.

C0208: **Exploring political narratives in European elections and their role for election success**
Presenter: **Ivan Savin**, ESCP Business School, Madrid campus, Spain
Co-authors: Jessica Birkholz, Peter Winker

The role of political narratives for success in political elections is hard to overestimate. Political parties promising to combat, e.g., inflation, unemployment or climate change, can be more successful than their rivals if their message meets the concerns of their voters. Over the last fifty years, Europe has experienced several crucial changes, ranging from the fall of the Iron Curtain and the global economic recession that started in 2007 through the strengthening of the European Union as a political and economic union to military conflicts in Yugoslavia and Ukraine. We plan to quantify the role of political narratives in fostering election success in European countries by considering country-specificities (e.g., their political system and history), their macroeconomic development, exposure to climate shocks, and the time of the election. We expect to build a map of political narratives present in European elections in the last half a century and demonstrate how different narratives were leading to better election results in different time periods and in different groups of countries (e.g. Western vs Eastern Europe, North vs South Europe). Quantifying the presence of different narratives in political party programs over time is the main purpose of the topic modelling method. We expect topics to capture different narratives political parties address during elections (poverty, unemployment, climate change, globalisation).

CO093 Room 45 ADVANCES IN BAYESIAN DEEP LEARNING Chair: Nadja Klein

C0284: **Partially Bayesian neural networks for adversarial robustness**
Presenter: **Tim-Moritz Buendert**, Technische Universitat Dortmund, Germany
Co-authors: Nadja Klein

Neural network models have been shown to be inherently vulnerable to adversarial examples, which significantly hinders their deployment in safety-critical applications. As one countermeasure, Bayesian deep learning was previously proposed to improve the robustness of such models. Still, most neural networks are constructed deterministically due to their efficient training and lower computational costs compared to the fully Bayesian counterpart. To combine the best of the deterministic and Bayesian deep learning approach, it is proposed to use partially Bayesian neural networks (pBNNs) to increase model robustness against adversarial attacks. To this end, a pre-trained deterministic neural network model is employed, and only a single selected layer is treated in a Bayesian fashion. This, in turn, keeps the computational efforts fairly low while still yielding complete posterior distributions rather than only point estimates. First, it is shown theoretically that under certain assumptions, some of these models are asymptotically robust to gradient-based adversarial attacks. Experimental results support this idea, highlighting enhanced adversarial robustness compared to deterministic neural networks and other competing approaches. Beyond adversarial robustness, the efficacy of pBNNs is demonstrated in further applications where uncertainty quantification is crucial, such as out-of-distribution detection.

C0362: **Advances in Bayesian neural model selection**
Presenter: **Alexander Immer**, ETH Zurich, Switzerland

Choosing optimal hyperparameters for deep learning can be highly expensive due to trial-and-error procedures and required expertise. Conceptually, a Bayesian approach to hyperparameter selection could help overcome such issues because it can rely on gradient-based optimization and does not require a held-out validation set. However, such an approach requires estimation and differentiation of the marginal likelihood, which is inherently intractable. Recent advances in Laplace approximations are discussed, which provide efficient estimates and enable optimizing hyperparameters with stochastic gradients just like neural network weights. Further, successful applications of Bayesian model selection are demonstrated, and shortcomings of current algorithms are discussed.

C0367: **Improving motion prediction in autonomous driving with expert knowledge: a Bayesian deep learning approach**
Presenter: **Christian Schlauch**, Humboldt Universitaet zu Berlin, Germany

Autonomous driving is one of the most highly anticipated yet elusive mobility innovations. The field has made significant advances through deep learning, especially in perception and motion prediction. Still, the field faces open challenges since safety requirements in autonomous driving demand robust domain adaptations between locations and well-calibrated uncertainty estimates for the numerous high-risk edge cases in urban environments. To meet those demands, the potential of Bayesian deep learning methods is explored for motion prediction. It demonstrates how a Bayesian approach can be used to regularize commonly employed motion prediction models by utilizing prior expert knowledge. More specifically, a CoverNet baseline model is adopted with a compute-efficient last-layer Gaussian Process approximation, and prior drivability knowledge is integrated. Doing so improves both robustness and calibration, as evaluated on various datasets.

C0384: **Provable guarantees for Bayesian neural networks**
Presenter: **Matthew Wicker**, Imperial College London, United Kingdom

To achieve the significant and substantial potential of modern machine learning in deployment, models must be optimized for both performance and reliability. It is no secret that deterministic neural networks suffer from diverse, critical failure modes, making their deployment challenging and requiring practitioners to formally verify that their NNs are safe for deployment. On the other hand, Bayesian neural networks (BNNs) appear to have favourable trustworthiness properties, including heightened robustness and fairness. In addition, BNNs offer many other considerable theoretical advantages. However, practical or theoretical advantages cannot be realized without meeting the bar of formal certification required of deterministic neural networks. Studies which provide empirical evidence for the heightened trustworthiness of BNNs are reviewed. Then, the broad spectrum of properties of interest for verification is covered when given a BNN, which will provide an intuitive methodology for computing formal guarantees for such properties. It is shown how to incorporate these properties into the likelihood at inference time to infer BNN posteriors with provably trustworthy predictions. Finally, the limitations of the methods presented are concluded with, and a brief but systematic coverage of some significant open research questions in the area are provided.

CC143 Room 001 APPLIED STATISTICS AND ECONOMETRICS Chair: Marialuisa Restaino

C0425: **Clustering the impact: How economic realities and political institutions shaped COVID-19 fiscal responses in Africa**
Presenter: **Samantha Joy Cinco**, Hochschule Fulda / University of the Free State, Germany

The fiscal response of African countries to the COVID-19 crisis are analyzed with an emphasis on how their responses varied based on their economic situations and political institutions prior to the start of the pandemic. A dataset of political and economic indicators prior to the pandemic (2019) and the total amount of fiscal spending during the pandemic (2020-2021) for all countries in Africa was leveraged. Using this data, OLS regressions were conducted to determine the most influential political and economic factors affecting fiscal response during the pandemic. These factors were then used in a K-means clustering approach to categorize African countries based on similar economic and political profiles. Upon the completion of the clustering, further testing was conducted to evaluate the significance of the clusters on the diverse fiscal response. Country clusters were determined using estimates of current account balance, government effectiveness, and political stability, controlled for the total number of reported COVID-19 cases. Results indicate that countries within the same cluster exhibit commonalities in their fiscal response and in their economic and political profiles. Moreover, subsequent Kruskal-Wallis and Dunns test results highlight the significance of these clusters, showing that economic context and political institutions influenced a country's approach to COVID-19, specifically in fiscal policy.

C0511: **Use of a nonparametric Bayesian method to model health state preferences: An application to Lebanese SF-6D valuations**
Presenter: **Samer Kharroubi**, American University of Beirut, Lebanon

This paper reports on the findings from applying a new approach to modelling health state valuation data. The approach applies a nonparametric model to estimate SF-6D health state utility values using Bayesian methods. The data set is the Lebanon SF-6D valuation study where a sample of 249 states defined by the SF-6D was valued by a representative sample of 577 members of the Lebanese general population using standard gamble. The paper presents the results from applying the nonparametric model and comparing it to the original model estimated using a conventional parametric random effects model. The covariates effect on health state valuations was also reported. The two models are compared theoretically and in terms of empirical performance. The nonparametric Bayesian model is argued to be theoretically more flexible and produces better utility predictions from the SF-6D than previously used classical parametric model. In addition, the Bayesian model is more appropriate to account the covariates effect. Further research is encouraged.

C0508: **Benchmark dose profiles for bivariate exposures**
Presenter: **Tugba Akkaya-Hocagil**, Ankara University, Turkey
Co-authors: Louise Ryan, Richard Cook, Sandra Jacobson, Joseph Jacobson

The Benchmark Dose has become a popular tool for identifying an exposure level that is associated with a specified increased risk of an adverse health outcome. However, the classic approach does not apply in settings where the exposure of interest can be measured in several different dimensions. In the application that motivates our work, for example, it is thought that cognitive deficits associated with prenatal alcohol exposure depend not simply on average alcohol consumption during pregnancy, but on both the frequency of drinking days as well as the amount of alcohol consumed on each drinking day during pregnancy. In this work, we propose a flexible framework for benchmark analysis that allows a more nuanced assessment of risks associated with an exposure that can be measured in several dimensions. The method entails fitting a joint model for the effect of the exposures while adjusting for potential confounders via propensity scores. From this model we obtain a benchmark dose contour that relates the two continuous exposure variables to an outcome. Additionally, we use generalized additive models which yield a flexible dose-response surface without requiring specification of a parametric form for non-linear effects. We illustrate our method using data assembled from six U.S. cohort studies that measured maternal reports of alcohol use during pregnancy, along with longitudinal measurements of cognitive function in their offspring.

CC039 Room 050 BIOSTATISTICS Chair: Andreas Mayr

C0230: **Brain-Network mathematical modeling for neurodegenerative disease**
Presenter: **Maria Mannone**, National Research Council (CNR) and University of Potsdam, Italy
Co-authors: Norbert Marwan, Peppino Fazio, Patrizia Ribino

The human brain can be described as a multi-layer network. We can model neurological disease in terms of alterations of the brain network at different layers. Thus, we define an operator acting on connectivity matrices and altering the weights of the connections. In particular, we can conceptualize an operator K, that acts on a healthy brain and produces a pattern of change typical for each disease, or describes the time evolution of a diseased brain. Focusing on neurodegenerative diseases having age as a risk factor, we consider Parkinson's and Alzheimer-Perusini's disease. We applied our model to patients from the Parkinson's Progression Markers Initiative (PPMI) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) datasets. We computed matrix forms of the K-operator comparing healthy control and diseased brain networks, and gained insights into the disease evolution by computing the K-operator between brains from the baseline to different follow-ups. We compared our findings with the medical literature, confirming the relevance of our results. Finally, we propose a machine-learning model to predict the patients' disease evolution, using as the training set our findings on the K-operator for different patients.

C0251: **A competing risk model for disease-specific or net survival estimation**
Presenter: **Reuben Adatorwovor**, University of Kentucky, United States

Standard survival analysis methods for cause-specific survival face challenges when the cause of event information is missing or unreliable, especially in cancer registry data. To address this, we introduce a methodology for enhanced disease-specific and net survival estimation, incorporating general background population mortality data. The approach, implemented in the dMrs package, offers robust net or relative survival estimation without relying on cause-specific-event information. It accommodates diverse dependence structures and facilitates comparison of disease-specific survival under dependent competing risks. We demonstrated the effectiveness of our approach by analyzing datasets from French breast cancer and Slovenian colon cancer, resulting in improved accuracy in survival estimates and enabling robust cohort comparisons. The dMrs package provides a versatile and reliable tool for researchers, enhancing our understanding of disease-specific and net survival.

C0344: **A supervised NMF extension for integrating omics data**
Presenter: **Aurelie Mercadie**, INRAE, France
Co-authors: Eleonore Gravier, Gwendal Josse, Nathalie Vialaneix, Celine Brouard

The motivation is to tackle a frequent problem in clinical research: patients stratified into K groups of interest (typically healthy/sick or control/treated patients) are described by biological measurements corresponding to different omics (metabolomics, proteomics, etc.). The aim is then to discover molecular signatures characterizing the groups. A non-negative matrix factorization (NMF) approach is extended to this framework. More specifically, this proposal is related to a supervised NMF, the FR-lda, which is based on an objective function that includes a supervised term to explain the groups of individuals. The proposal adapts this method to a multi-table framework by integrating information through a contribution matrix common to all omics. Compared to FR-lda, the supervised term is reworked to account for the non-negativity of the solution and resumes a criterion equivalent to K independent linear regressions, one for each group. Finally, two optimization methods are proposed to solve the induced optimization problem: the classical multiplicative approach (MU) and a novel proximal approach that achieves exact sparsity in molecular signatures, easing result interpretation. The method has been successfully tested on simulated as well as real data and compared with state-of-the-art methods for omics integration.

C0474: **Comparison of joint species distribution models for percent cover data**

Presenter: **Pekka Korhonen**, University of Jyvaskyla, Finland

Co-authors: Francis Hui, Sara Taskinen, Jenni Niku, Bert van der Veen

Joint species distribution models (JSDMs) have gained considerable traction among ecologists over the past decade due to their capacity to answer a wide range of questions at both the species- and the community level. The family of generalized linear latent variable models, in particular, has proven popular for building JSDMs, being able to handle many response types, including presence-absence data, biomass, overdispersed and/or zero-inflated counts. Latent variable models are extended to handle per cent cover data, with vegetation, sessile invertebrate, and macroalgal cover data representing the prime examples of such data arising in community ecology. Sparsity is a commonly encountered challenge with per cent cover data. Responses are typically recorded as percentages covered per plot, though some species may be completely absent or present, i.e., have 0% or 100% cover, respectively, rendering the use of beta distribution inadequate. Two JSDMs are proposed suitable for per cent cover data, namely a hurdle beta model and an ordered beta model. The two proposed approaches are compared to a beta distribution for shifted responses, transformed presence-absence data, and an ordinal model for per cent cover classes. Results demonstrate the hurdle beta JSDM was generally the most accurate at retrieving the latent variables and predicting ecological per cent cover data.

Thursday 29.08.2024 16:00 - 17:30 Parallel Session K – COMPSTAT2024

CV052 Room 001 COMPLEX DATA ANALYSIS Chair: Matthieu Marbac

C0420: **Functional motif discovery in stock market prices**
Presenter: **Federico Severino**, Universite Laval, Canada
Co-authors: Marzia Cremona, Lyubov Doroshenko

Financial asset prices display recurrent patterns over time. However, such time series are usually noisy and volatile, making the identification of repetitive patterns challenging. These motifs are rarely exploited for price prediction, even though some of them, such as the surge of a financial bubble, occur periodically and feature similar shapes. Asset prices are embedded in a functional data analysis framework by extending probabilistic K-means with local alignment to discover functional motifs in stock price time series. The discovered motifs are then exploited for price forecasting by developing a novel motif-based (MB) algorithm. After illustrating the technique on simulations of mixed causal-noncausal autoregressive processes, it is applied to the prices of S&P 500 top components. Finally, the superior performance of motif-based forecasting is demonstrated in several other forecasting methods.

C0467: **The MEM algorithm and modal clustering of functional data**
Presenter: **Adhiraj Mandal**, University of Glasgow, United Kingdom

Functional data analysis (popularly abbreviated as FDA) is a branch of statistics that attempts to analyse information on a process that varies over a continuum. Such processes are often considered to be functions of time, though they can also be any other domain, such as energy, spatial location, wavelength, etc. In the FDA, each curve is considered to be an individual entity instead of a number of individual observations along the curve. Though this is a rich source of information about the process of generating the data, it also makes both theoretical and computational work centered on functional data challenging. The main focus of my research is to develop a framework for modal clustering for clustering functional data. The discussion starts with the MEM (modal expectation maximization) algorithm for multivariate data on a finite-dimensional space, followed by the HMAC (hierarchical mode association clustering) and PHMAC (parallel HMAC) algorithms for clustering functional data. The MEM algorithm plays a crucial role in the functionality and effectiveness of the HMAC and PHMAC algorithms. The benefits and drawbacks of the algorithms are discussed, and how these techniques might be applied for clustering functional data is investigated. In order to observe the clustering outcomes, the identical algorithms are applied to two data sets: one simulated and one real-world.

C0482: **Outlier detection in mixed data**
Presenter: **Houda Gadacha**, CNAM Paris, France
Co-authors: Patricia Kubicki, Ndeye Niang

Outlier detection is crucial in various fields, such as insurance fraud, disease detection, and cybersecurity. Its application helps to identify suspicious behaviors and enhance the robustness of statistical models. Most outlier detection methods are designed exclusively for numerical data. To detect outliers in data containing both numerical and categorical attributes, factor analysis for mixed data (FAMD) is proposed to extract numerical components. These components are then used for outlier detection. Outlier detection methods are applied only to the first components, the last components, and all FAMD components. The results are compared to those of a traditional one-hot encoding (OHE) preprocessing approach based on simulated data. The simulated data includes four outlier types: (a) global outliers, which significantly deviate from most data points; (b) local outliers, which are not necessarily extreme values but are considered abnormal within their specific context or neighborhood, (c) rare outliers which have unexpected categories compared to the typical data distribution, and (d) mixed outliers which can be both global and rare, or local and rare. The objective is to determine the most effective method in terms of outlier types detected. The results demonstrate the effectiveness of the proposed approach.

C0483: **Different approaches for modeling multivariate space-time data: A performance-based comparison**
Presenter: **Claudia Cappello**, University of Salento, Italy
Co-authors: Sandra De Iaco, Monica Palma, Klaus Nordhausen

In the last decades, great advances have been made in the multivariate space-time framework for modeling the matrix-valued covariance function. One of the first models proposed in the literature has been the space-time linear coregionalization model (ST-LCM), developed for modeling matrix-valued covariance function as a linear combination of univariate models related to latent uncorrelated processes underlying the investigated phenomenon. Moreover, in recent years, the space-time blind source separation-based model (ST-BSS) has been proposed as an alternative modeling approach based on the univariate analysis of the latent components. In the application stage, each of the above-mentioned techniques is characterized by its own drawbacks and advantages in terms of model parameters to be estimated and/or time-consuming in the modeling procedure. The different modelling steps and the predictive performances of the ST-LCM and the ST-BSS model are focused from a both theoretical and practical points of view, and a comparison of their performances is provided through a simulation study and an application on a space-time environmental data set.

CV025 Room 050 APPLIED STATISTICS AND EMPIRICAL ANALYSES Chair: Enea Bongiorno

C0500: **Some new insights into measures of location and adopting voting technique into measures of variability**
Presenter: **Kayode Ayinde**, Northwest Missouri State University, Maryville, Missouri, USA, United States

Both measures of location and variability in data sets have become increasingly popular and relevant in almost all fields. There arises a question of how agreeable or acceptable each measure of location is to the individual subjects, as this may not be the most representative. Some new insights into measures of locations are considered by incorporating the existing measures into the five and three summary statistics of a data set. The voting technique is also adopted as a measure of variability to address the challenge of acceptance by most subjects. Monte Carlo simulation studies of symmetric and skewed data sets with and without outlier(s) were conducted on the measures, and real-life data sets were applied. Results reveal the consistent minimization of the mean absolute deviation by the median and mean squared deviations by the arithmetic mean and that the challenge of non-existence and non-uniqueness of mode can be overcome with the voting technique as any of the averages can emerge as the most representative average. Furthermore, the harmonic mean of the five summary statistics is identified as the best average, especially with data sets that have outliers in the right direction in both simulation and real-life application studies.

C0401: **Classical, Bayesian, and machine learning variable selection methods in colorectal cancer microbiome study: A comparison**
Presenter: **Mohammad Fayaz**, Allameh Tabatabaei University, Iran
Co-authors: Edda Russo , Leandro Di Gloria, Sara Bertorello, Amedeo Amedei

Human microbiome research requires new and appropriate statistical methods to analyze the complex structure of microbiome datasets. A previously published microbiome dataset based on 46 Colorectal Cancer (CRC) and 15 Adenomatous Polyps (AP) patients are utilized, along with their microbiome samples from three sampling sites: saliva, tissue, and stools. The microbiome composition at five taxonomic levels (Phylum, Class, Order, Family, and Genus) was separately analyzed. The statistical challenge lies in classifying between CRC and AP and selecting variables for metagenomic features. Initially, classical methods such as LASSO, RIDGE regression, zero-inflated beta regressions (ZIB), generalized additive models for location scale and shape (GAMLSS-BEZI), and compositional data analysis (CoDA) were compared using the SIAMCAT etc. libraries in R. Subsequently, Bayesian model averaging (BMA) methods for generalized linear models (GLM) with different priors were employed using the

BMA and BAS packages. Additionally, various interpretable machine learning (IML) algorithms, including variable importance plot (VIP), partial dependence plot (PDP), local interpretable model-agnostic explanations (LIME), and Shapley values for machine learning methods (ML) such as random forests, were explored. Finally, a concordance plot of selected metagenomics across different methods is presented.

C0459: **Comparing curves with the discrete Frechet distance via the family of exponentiated generalized distributions**
Presenter: **Mireya Diaz**, Case Western Reserve University, United States

The discrete Frechet distance is widely used in applications assessing the similarity between curves in areas including pattern recognition, map routing, protein structure alignment, and time series clustering. It is defined as the minimum among the maxima of a series of feasible paths along the curves under comparison connecting the first and last differences between these curves. Despite its use in diverse problems and profuse computational developments, little is known about its distributional properties and, thus, its power in statistical inference. A simulation study allowed the assessment of such distributional properties empirically under null and alternative scenarios for four families of functions: linear, quadratic, sinusoidal, and exponential. A parallel analytic work using extreme value theory under dependence identified its distribution. The discrete Frechet distance belongs to the family of exponentiated generalized distributions. One of these corresponds to the Kumaraswamy distribution, a beta-type distribution. Under conditions of exchangeability of the random variables, the distribution of the discrete Frechet distance is bounded by the conventional cumulative distribution function of the minimum. The empirical side confirmed this by the 2- and 4-parameter beta distributions, which fit most of the histograms adequately for both null and alternative scenarios. These findings allow the application of the Frechet distance under an inferential framework.

CV057 Room 051 COMPUTATIONAL STATISTICS AND APPLICATIONS Chair: Thomas Kneib

C0280: **A stochastic expectation maximisation approach to record linkage**
Presenter: **Kayane Robach**, Amsterdam UMC, Netherlands
Co-authors: Stephanie van der Pas, Mark van de Wiel, Michel Hof

A new method is introduced to combine observations from overlapping data sets without a unique identifier, commonly known as record linkage. This task holds potential importance in healthcare longitudinal studies where one has to rely on partial information to monitor the data, and more broadly, it offers the opportunity to augment data with external sources, circumventing costly data collection. As the main innovations, we address time-varying variables like place of residence and develop an efficient algorithm that can be used on large data sources. The complexity of the record linkage task stems from the sub-par reliability of the partially identifying variables (e.g. initials, birth year, zip code) used to link records and their limited number of unique values. Furthermore, because everyone is often uniquely represented in each file, records from one file can maximally be linked with one record in the other file, making the linkage decisions interdependent. Our new approach uses a Stochastic Expectation Maximisation based on a latent variable model to accommodate registration errors and changes in the identifying information over time. Our model provides a probabilistic estimate of the common set of records that allows for inference. We implement our methodology in an R package, investigate its properties with a simulation study, and apply it to two large surveys.

C0414: **Copula approximate Bayesian computation using distribution random forests**
Presenter: **George Karabatsos**, University of Illinois-Chicago, United States

A novel approximate Bayesian computation (ABC) framework is introduced for estimating the posterior distribution and the maximum likelihood estimate (MLE) of the parameters of models defined by intractable likelihood functions. This framework can describe the possibly skewed and high dimensional posterior distribution by a novel multivariate copula-based distribution based on univariate marginal posterior distributions, which can account for skewness and be accurately estimated by distribution random forests (DRF) while performing automatic summary statistics (covariates) selection, and on robustly estimated copula dependence parameters. The framework employs a novel multivariate mode estimator to perform MLE and posterior mode estimation and provides an optional step to perform model selection from a given set of models with posterior probabilities estimated by DRF. The posterior distribution estimation accuracy of the ABC framework is illustrated through simulation studies involving models with analytically computable posterior distributions and exponential random graph and mechanistic network models, each defined by an intractable likelihood from which it is costly to simulate large network datasets. Also, the framework is illustrated through analyses of large real-life networks of sizes ranging between 28,000 to 65.6 million nodes (between 3 million to 1.8 billion edges), including a large multilayer network with weighted, directed edges.

C0217: **R package QuantileGH: Quantile least Mahalanobis distance estimator for Tukey g-&-h mixture**
Presenter: **Tingting Zhan**, Thomas Jefferson University, United States
Co-authors: Misung Yi, Inna Chervoneva

A mixture of 4-parameter Tukey g-&-h distributions is proposed for fitting finite mixtures with Gaussian and non-Gaussian components. Since the likelihood of Tukey's g-&-h mixtures does not have a closed analytical form, we propose a Quantile Least Mahalanobis Distance (QLMD) estimator for the parameters of such mixtures. QLMD is an indirect estimator minimizing the Mahalanobis distance between the sample and model-based quantiles, and its asymptotic properties follow from the general theory of indirect estimation. We have developed a stepwise algorithm to select a parsimonious Tukey g-&-h mixture model and implemented all proposed methods in the R package QuantileGH available CRAN. A simulation study was conducted to evaluate the performance of the Tukey g-&-h mixtures and compare them to the performance of mixtures of skew-normal or skew-t distributions. The Tukey g-&-h mixtures were applied to model cellular expressions of Cyclin D1 protein in breast cancer tissues, and resulting parameter estimates were evaluated as predictors of progression-free survival.

C0185: **Highly robust training of regularized radial basis function networks**
Presenter: **Jan Kalina**, The Czech Academy of Sciences, Institute of Information Theory and Automation, Czech Republic

Radial basis function (RBF) networks represent established tools for nonlinear regression modeling with numerous applications in various fields. Because their standard training is vulnerable with respect to the presence of outliers in the data, several robust methods for RBF network training have been proposed recently. The focus is on robust regularized RBF networks. A robust inter-quantile version of RBF networks based on trimmed least squares is proposed. Then, a systematic comparison of robust regularized RBF networks follows, which is evaluated over a set of 405 networks trained using various combinations of robustness and regularization types. The experiments proceed with a particular focus on the effect of variable selection, which is performed by means of a backward procedure, on the optimal number of RBF units. The regularized inter-quantile RBF networks based on trimmed least squares turn out to outperform the competing approaches in the experiments if a highly robust prediction error measure is considered.

CO107 Room 052 COMPUTATIONAL METHODS FOR STATISTICAL LEARNING Chair: Mauro Bernardi

C0281: **State-space models for clustering of compositional trajectories**
Presenter: **Andrea Panarotto**, Department of Statistical Sciences, University of Padova, Italy
Co-authors: Manuela Cattelan, Ruggero Bellio

Compositional data are drawing increasing interest for their ability to depict interdependent and constrained observations. While time series analysis has sometimes been employed to study individual compositional trajectories, little attention has been given to finding and modeling groups of trajectories. Driven by a sustainable mobility motivation, we propose a model-based approach, relying on a state space model representation and an Expectation-Maximization algorithm, for clustering compositional trajectories according to their evolution in the simplex. Trajectory covariates not captured by the compositional representation can be included in the group assignment phase of the method. The model is applied to urban movement data, where people's movements are represented in the simplex by the proportions of road types in their surroundings.

C0340: **Learning algorithms for constrained hidden Markov models**
Presenter: **Connor Mattes**, Sandia National Laboratory, United States
Co-authors: William Hart

Hidden Markov models (HMMs) are an invaluable tool for modeling discrete processes with hidden (latent) variables. HMMs have proven useful in fields ranging from speech recognition and natural language processing to bioinformatics and bird migration. The focus is specifically on constrained HMMs. Constrained HMMs capture subject matter expertise by restricting the set of hidden variables. For example, hidden states may incur costs that must be bounded, or it may require that some hidden state occur at least once. Although there has been some prior research on learning for constrained HMMs there are significant gaps in the literature. Learning in constrained HMMs is difficult because it is computationally intractable to sum over all feasible hidden states. Towards this end, three novel learning algorithms are introduced that efficiently approximate model parameters by generating many high-quality feasible sets of hidden states. The advantages and disadvantages of each of these new algorithms are discussed. Those are compared theoretically and computationally to the existing literature.

C0393: **Dynamical quantile graphical modeling**
Presenter: **Piergiacomo Andrea Carlesi**, Univ. degli Studi di Padova, Dept. of Statistical Science, Italy
Co-authors: Mauro Bernardi, Cristian Castiglione, Nicolas Bianco

An innovative approach is introduced for jointly estimating multiple quantiles within a dynamic framework. Essentially, the method acknowledges that the fixed-level quantile of a vector of response variables is influenced by both a fixed set of covariates and a random effect with an autoregressive dynamic. As a result, the proposed framework expands upon traditional univariate quantile models by integrating a vector autoregressive structure. To streamline the estimation of model parameters and the extraction of signals, Bayesian methodologies are developed that leverage approximate techniques and data augmentation strategies. These methodological advancements are instrumental in efficiently addressing the complexities inherent in estimating the model parameters and extracting meaningful signals from the data. A comprehensive panel comprising US equity market returns and macroeconomic indicators is utilized to empirically validate our approach. Through this analysis, the objective is to shed light on the dynamic evolution of spillover effects within individual Value-at-Risk estimates. By scrutinizing the interaction between macroeconomic variables and the autoregressive component, the aim is to uncover the intricate mechanisms that drive risk dynamics in financial markets.

C0346: **Dynamic network models with time-varying nodes**
Presenter: **Luca Gherardini**, University of Florence, Italy

Networks are used to represent complex data structures that arise in different fields of science and are often not static objects, as they can evolve over time. The main approaches in the literature belong to the broad class of latent variable models, including latent space models and stochastic block models. However, most methods were developed to model the dynamic behaviour of edges without considering that the network's topology may vary over time. It is shown that ignoring this new source of complexity can lead to non-negligible bias in the parameter estimates since the model cannot discriminate structural zeros due to the topology of the network from those related to the absence of an edge between a pair of nodes. A class of zero-inflated Bernoulli model is proposed that embeds the time-varying node probabilities into the edge process, with the aim of modelling the entire dynamic network over time. A fully latent approach is also proposed for the nodes process, providing a probabilistic characterization for any network model (static and dynamic) that explicitly assumes a two-layer hierarchy, one for nodes and one for edges, regardless of the modelling choice. The inference approach for this class of models is developed within the Bayesian paradigm and relies on a Gibbs sampling algorithm with a Polya-gamma data augmentation scheme. The performance of the approach is explored through a simulation study.

CO122 Room 43 ADVANCES IN FINITE MIXTURES FOR REGRESSION AND CLUSTERING Chair: Gabriele Soffritti

C0279: **Full model estimation for non-parametric multivariate finite mixture models**
Presenter: **Matthieu Marbac**, CREST - ENSAI, France
Co-authors: Marie du Roy de Chaumaray

The problem of full-model estimation for non-parametric finite mixture models is addressed. We present an approach for selecting the number of components and the subset of discriminative variables (i.e. the subset of variables having different distributions among the mixture components) by considering an upper bound on the number of components (this number being allowed to increase with the sample size). The proposed approach considers a discretization of each variable into B bins and a penalization of the resulting log-likelihood. Considering that the number of bins tends to infinity as the sample size tends to infinity, we prove that our estimator of the model (number of components and subset of relevant variables for clustering) is consistent under a suitable choice of the penalty term. The relevance of our proposal is illustrated on simulated and benchmark data.

C0207: **Parsimonious mixtures of dimension-wise scaled normal mixtures**
Presenter: **Luca Bagnato**, Catholic University of the Sacred Heart, Italy
Co-authors: Antonio Punzo, Salvatore Daniele Tomarchio

A new family of parsimonious mixture models is introduced for model-based clustering. Dimension-wise scaled normal mixtures (DSNMs), recently introduced in the literature, are considered as mixture components. DSNMs generalize the multivariate normal (MN) distribution in two directions. Firstly, they have a more general type of symmetry with respect to the elliptical symmetry of the MN distribution. Secondly, the univariate marginals have similar heavy-tailed normal scale mixture distributions with (possibly) different tailedness parameters; as a consequence of practical interest, DSNMs allow for a different excess kurtosis on each dimension. Due to the structure of these mixture components, parsimony is attained through the variance-correlation decomposition. A variant of the expectation-maximization algorithm is presented for maximum likelihood parameter estimation. Parameter recovery and clustering performance are investigated via a simulation study. Comparisons with the unconstrained mixture models are obtained as by-products. Lastly, our and the competing models are evaluated in terms of fitting and clustering on some real datasets.

C0295: **Mixtures of skewed regression models for clustering spatial data**
Presenter: **Michael Gallaugher**, Baylor University, United States

Over the years, data has become increasingly complex, creating the need for new analytical tools. This is especially the case in the areas of clustering and classification, as well as spatial analysis. In the case of a large and complex spatial domain with predictor-response relationships between variables, a single regression model is unlikely to hold. To address this challenge, we propose a mixture of regression models on a Markov random field combined with skewed distributions for clustering spatial data. The model identifies clusters of locations with similar predictor-response relationships. Overfitting in the number of groups is addressed by integrating skewed distributions into the error term of each component of the mixture of regressions. Parameter estimation via an EM approach will be discussed and assessed via simulation. Insurance data and water basin data will be used for illustration.

C0361: **Sparse multivariate Gaussian mixture regression with covariance estimation**
Presenter: **Gabriele Soffritti**, University of Bologna, Italy
Co-authors: Michael Fop, Marco Vitelli

Gaussian clusterwise regression models and Gaussian cluster-weighted models represent useful tools to simultaneously perform multivariate linear regression analysis and model-based cluster analysis in the presence of continuous variables when the population from which the sample comes is composed of a certain number of sub-populations and the specific sub-population each sample observation belongs to is unknown. As the number of parameters scales quadratically with the number of variables, such models can be over-parameterized in the case of high-dimensional data. To mitigate this problem, lasso penalties are introduced in the model log-likelihood function so as to simultaneously obtain sparse estimators of the regression coefficients and the covariance structure within each component of the mixture. For this purpose, an efficient optimization procedure is embedded into the expectation-maximization algorithm usually employed to perform maximum likelihood estimation. The new penalized Gaussian clusterwise linear regression models and Gaussian cluster-weighted models obtained in this way allow both variable selection and regularization to be performed. This approach is also expected to enhance the prediction accuracy and interpretability of regression analysis and cluster analysis based on Gaussian clusterwise regression models and Gaussian cluster-weighted models. The performance of the new methodology is studied through simulated and real data.

CO092 Room 44 HiTEc: High-dimensionality and time series Chair: Andreas Artemiou

C0298: **Flexible extreme marginal quantile treatment effect in high dimensions**
Presenter: **Jing Zhou**, University of East Anglia, United Kingdom

The marginal quantile treatment effects are investigated when the quantile level approaches 0 or 1. When the quantile level approaches the ends, quantile regression cannot accurately model the tail distributions. To overcome this limitation, we propose an alternative approach that uses extreme quantile models to estimate the marginal effect in the presence of a continuous covariate shift. Such models use an extreme value index to model the tail of the distribution function. This method estimates an extreme value index at intermediate quantile levels and extrapolates to the tails where the quantile level is close to zero. By extrapolating, we aim to estimate the extreme treatment effects consistently and obtain the corresponding asymptotic distribution. Further, when the number of parameters is nonnegligible, we consider regularization to identify the relevant covariates among hundreds of variables for the extreme quantile treatment effects.

C0250: **Fully modified estimation of a quantile cointegration model with a spatial lag**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Christian Haefke

A quantile cointegration regression problem is investigated, which includes a spatial lag. We obtain the asymptotic distribution of the quasi-maximum likelihood estimator and show that the second-order bias depends on the order of the deterministic terms. We construct both a fully modified estimator, which removes this second-order bias and a Wald-type test. The finite sample properties of our fully modified estimator are analyzed in a simulation study. Finally, our estimation procedure is used to analyze growth-at-risk for a set of countries, where the cross-country spillover effects are described by means of a spatial lag.

C0260: **Computationally efficient SVM-based sufficient dimension reduction**
Presenter: **Andreas Artemiou**, University of Limassol, Cyprus

Support-vector-machine-based (SVM-based) sufficient dimension reduction has been shown to be a great alternative to the inverse moment-based methodology introduced in the early 90s because it allows for linear and nonlinear feature extraction under a unified framework. Although it performed a more accurate estimation of the dimension reduction subspace of interest (also known as the Central Subspace), it was computationally more challenging. Although a number of projects have addressed this issue, we present an alternative approach which utilizes Twin-SVM.

C0477: **Order determination in second-order source separation models using data augmentation**
Presenter: **Klaus Nordhausen**, University of Jyvaskyla, Finland
Co-authors: Una Radojicic

A robust estimator is proposed for determining the number of latent components in an internal noise model within the second-order source separation (SOS) framework. The method incorporates a data augmentation strategy along with the robust SOS approach, eSAM-AMUSE, which utilizes information from eigenvalues and variations of eigenvectors. The dimension estimate derived from the approach can be visualized using a ladle plot. Through a simulation study, the new estimator is demonstrated to exhibit superior properties and consistently outperform the bootstrap-based AMUSEladle estimator.

CO100 Room 45 Advances in multivariate distributional regression Chair: Guillermo Briseno Sanchez

C0235: **Handling endogenous regressors in quantile regression models: Copula approach without instruments**
Presenter: **Rouven Haschka**, Zeppelin University Friedrichshafen, Germany

Endogeneity in quantile regression models has not yet received much attention in the literature. In order to handle regressor endogeneity, only instrument-based approaches are used. Seeing that instruments are often weak or unavailable, this article proposes a generalisation for the instrument-free Gaussian copula-based endogeneity correction to quantile regression models. For this purpose, we develop two estimators. The first is a full maximum likelihood estimator based on directly maximising the likelihood derived from the joint distribution of explanatory variables and errors, given the assumption that errors follow an asymmetric Laplace distribution. The second is a Bayesian estimator based on a decomposition of errors into an unobserved exponential variable and a structural normal part. By assuming that endogeneity comes from the normally distributed part, we can use the copula endogeneity correction by control functions so that the model can be estimated by efficient Gibbs sampling. We derive identification assumptions and show under which conditions these are fulfilled. Moreover, we provide Monte Carlo simulation results to examine and compare the finite sample performances of the two abovementioned estimators and demonstrate their superiority to instrumental variable estimation in quantile regression models.

C0239: **Deep mixture of linear mixed models for complex longitudinal data**
Presenter: **Lucas Kock**, National University of Singapore, Singapore
Co-authors: Nadja Klein, David Nott

Mixtures of linear mixed models are widely used for modeling longitudinal data for which observation times differ between subjects. In typical applications, temporal trends are described using a basis expansion, with basis coefficients treated as random effects varying by subject. A key advantage of these models is that they provide a natural mechanism for clustering, which can be helpful for interpretation. Current versions of mixtures of linear mixed models are not specifically designed for cases where there are many observations per subject and a complex temporal trend, which requires a large number of basis functions to capture. In this case, the subject-specific basis coefficients are a high-dimensional random effects vector, for which the covariance matrix is hard to specify and estimate, especially if it varies between mixture components. To address this issue, we consider the use of recently developed deep mixture of factor analyzers models as the prior for the random effects. The resulting deep mixture of linear mixed models is well-suited to high-dimensional settings. We demonstrate the adaptability of our approach across a range of real-world applications, including within-subject prediction for an unbalanced longitudinal study, the task of predictive likelihood-free inference as well as missing data imputation.

C0318: **Generalized additive models for smoothing covariance matrices**
Presenter: **Vincenzo Gioia**, University of Trieste, Italy
Co-authors: Matteo Fasiolo, Ruggero Bellio

Computational and methodological developments in multi-parameter Generalized Additive Models (GAMs), also known as distributional regression models, have been the primary drivers of their popularity in applied statistics. The multivariate Gaussian additive model, where both the elements of the mean vector and an unconstrained parametrisation of the covariance matrix are modelled semi-parametrically, exemplifies this class of models. The task of modelling the covariance matrix is complex, and ensuring scalability during model fitting is challenging, primarily due to the high dimensionality of the problem and the necessity of performing smoothing parameter optimisation. Furthermore, the advantage of adopting an unconstrained parametrisation of the covariance matrix complicates the interpretation of the model output, which needs to be translated into well-known quantities such as variance and correlation. In the proposed approach, the computational challenge is addressed by adopting efficient computational strategies, which are integrated into well-established GAMs' model fitting routines, while interpretability issues are mitigated by utilising accumulated local effects. The proposed modelling approach is illustrated with real-world examples, with particular emphasis on applications in the energy sector.

C0339: **A model-based boosting approach to deal with dependent censoring**

Presenter: **Annika Stroemer**, University of Bonn, Germany

Co-authors: Nadja Klein, Andreas Mayr

A popular model to study the effect of covariates on survival times is the semiparametric proportional hazards model. Estimation in this model is well-established for common right-censored data, assuming independence between survival and censoring time given the covariates. This assumption is mainly held when censoring occurs at the end of the study. However, in medical studies, this assumption is questionable. For example, if a patient's health deteriorates and they choose to withdraw from the trial due to poor prognosis, censoring time depends on their health status. This leads to dependent censoring, as patients with poorer health are more likely to be censored earlier. A model-based boosting approach is proposed to deal with dependent censoring using distributional copula regression. This allows modeling the joint distribution of survival and censoring times by linking appropriately marginal distributions through a parametric copula. Rather than assuming known marginals, all distribution parameters are estimated simultaneously as functions of covariates. A key merit of the boosting approach compared to classical estimation frameworks is that estimation is feasible for high-dimensional data. Additionally, the boosting algorithm includes data-driven variable selection. An extensive simulation study is conducted, and its practical application is illustrated with a biomedical example.

Friday 30.08.2024 09:00 - 10:30 Parallel Session M – COMPSTAT2024

CO095 Room 052 MARKETING AND INNOVATION Chair: Yuichi Mori

C0231: **Graphical copula GARCH modelling with dynamic conditional dependence**
Presenter: **Mike So**, The Hong Kong University of Science and Technology, Hong Kong
Co-authors: Shun Hin Chan, Amanda Chu

Traditional correlation models, such as the dynamic conditional correlation (DCC)-GARCH model, often omit the nonlinear dependencies in the tails. We aim to develop a framework to model the nonlinear dependencies dynamically among a large portfolio of stocks, namely the graphical copula GARCH (GC-GARCH) model. Motivated by the capital asset pricing model to allow high-dimensional modelling for large portfolios, the number of parameters can be greatly reduced by introducing conditional independence among stocks given the risk factors, such as the Hang Seng index in Hong Kong. The joint distribution of the risk factors is factorized using a directed acyclic graph (DAG) with pair-copula construction (PCC) to introduce flexibility. The DAG induces topological orders to the risk factors, which can be regarded as a list of directions of the flow of information. The conditional distributions among stock returns are also modeled using PCC. Dynamic conditional dependence structures are also incorporated to allow the parameters in the copulas to be time-varying. Three-stage estimation is used to estimate parameters in the marginal distributions, the risk factor copulas, and the stock copulas. In the investment experiments in the empirical study, we show that the GC-GARCH model produces more accurate conditional value-at-risk predictions and much higher cumulative portfolio values than the DCC-GARCH model.

C0377: **Analysis of consumer purchasing attitudes toward organically grown vegetables**
Presenter: **Yumi Asahi**, Tokyo University of Science, Japan

The production of vegetables in Japan has declined since the late 1980s because there were fewer agricultural workers. On the other hand, the number of imported vegetables has increased because the price of imported vegetables is low, and a stable supply of them is possible. However, in recent years, the consumers have become increasingly aware of problem related to the safety of food like chemical levels in imported vegetables, therefore, needs of domestic vegetable have risen. Therefore, the focus is on organically grown vegetables and search for the preference considerations that the consumer values when purchasing vegetables. To understand the causal relationship, analyze the factors related to buying organically grown vegetables. As a result of the analysis, the preference consideration on purchases of the vegetable became the result of valuing in order of price, growing area, cultivation method and production management. The purchasing factor of organically grown vegetables includes health trends, safety awareness, food concerns, regional society concerns, and environmental concerns. The development of social motivations has led to a change in attention to the entire food chain and has been linked to the purchase of organically grown vegetables.

C0355: **Applying machine learning approach to marketing uncovering consumer insights through big data**
Presenter: **Atsuho Nakayama**, Tokyo Metropolitan University, Japan

Vast amounts of marketing data are now available online. Automated collection of online clickstreams, messaging, word-of-mouth, transactional, location, and other data has significantly reduced the cost of data collection. The amount of data available today is increasing, and consumer behavior can be understood in great detail. In practice, the use of machine learning methods (including deep learning and cognitive systems) is encouraged. In recent years, convolutional neural networks (CNNs) have become the dominant algorithm for many computer vision tasks (acquisition, processing, and analysis of digital images), and many studies and applications using deep learning and AI have been conducted. The impact of them on marketing operations is expected to increase in the future. The question is how deep learning approaches should be used in marketing. Therefore, examples of applying machine learning and deep learning approaches are presented to the vast amount of consumer behavior data currently available, with the goal of supporting marketing decisions. The aim is to contribute to marketing research by deriving useful knowledge for market segmentation and positioning strategy formulation.

C0357: **Measuring technology acceptance over time by online customer reviews based transfer learning**
Presenter: **Daniel Baier**, University of Bayreuth, Germany
Co-authors: Andreas Karasenko, Alexandra Rese

Online customer reviews (OCRs) are user-generated semi-formal evaluations of objects (brands, companies, products, services, technologies). They typically consist of a time stamp, a star rating (1 to 5 stars) of the evaluated object and, in many cases, a natural language comment that details the perceived strengths and weaknesses. Up to now, many methodological approaches have been developed and applied to analyze and aggregate OCRs, as well as to improve products and services based on this knowledge. So, a prior study applied a lexicographic text mining approach similar to sentiment analysis to OCRs of IKEAs augmented reality app. They predicted construct scores for the extended technology acceptance model (TAM) and validated these predictions by conducting an additional extended TAM survey among app users. A new transformers-based approach is presented for the same purpose. A transfer learning model is trained, tested, and validated based on large samples of OCRs and corresponding extended TAM construct scores given by experts. The results are promising. They go beyond conducting an extended TAM survey for an object by validly predicting the development of construct scores over time.

CO097 Room 43 RECENT ADVANCES IN DESIGN AND ANALYSIS OF EXPERIMENTS Chair: Frederick Kin Hing Phoa

C0179: **Optimal exact designs for small studies in toxicology with applications to hormesis via metaheuristics**
Presenter: **Ray-Bing Chen**, National Cheng Kung University, Taiwan

There are theory-based methods for constructing model-based optimal designs when the sample size is large. The problem becomes challenging when the sample size is small. The theory may no longer apply, and even if it does, the optimal design may not be implementable. We provide examples and also show that a simple rounding procedure of the weights from an optimal approximate design to an optimal exact design can produce the wrong optimal exact design. To solve this longstanding, serious and practical problem, we propose a state-of-the-art nature-inspired metaheuristic algorithm to find efficient designs for an experiment with a small sample size. As an application, we use the algorithm to find an optimal design for a toxicology experiment to detect the existence of hormesis in a dose-response study and an optimal design to estimate the hormesis threshold. Being a metaheuristic algorithm, it can be used to find different types of optimal designs for various statistical models. We demonstrate its flexibility by finding locally D-optimal designs for estimating model parameters in logistic models for small experiments, along with user-friendly codes to produce all the designs.

C0180: **An efficient approach for identifying important biomarkers for biomedical diagnosis**
Presenter: **Frederick Kin Hing Phoa**, Academia Sinica, Taiwan
Co-authors: Jing-Wen Huang, Yan-Hong Chen, Yan-Han Lin, Shau Ping Lin

The challenges associated with biomarker identification for diagnosis purpose in biomedical experiments are explored. A novel approach is proposed to handle the above challenging scenario via the generalization of the Dantzig selector. To improve the efficiency of the regularization method, we introduce a transformation from an inherent nonlinear programming due to its nonlinear link function into a linear programming framework under a reasonable assumption on the logistic probability range. We illustrate the use of our method in an experiment with binary response, showing superior performance in biomarker identification studies when compared to their conventional analysis. The proposed method does not merely serve as a variable/biomarker selection tool; its ranking of variable importance provides valuable reference information for practitioners to reach informed decisions regarding the prioritization of factors for further investigations.

C0253: **Summary of effect aliasing structure for design selection and factor-column assignment for supersaturated designs**
Presenter: **Yi-Hua Liao**, National Tsing Hua University, Taiwan
Co-authors: Frederick Kin Hing Phoa, David Woods

In the assessment and selection of supersaturated designs, the aliasing structure of interaction effects is usually ignored in traditional criteria such as $E(s^2)$-optimality. We introduce the Summary of Effect Aliasing Structure (SEAS) to assess the aliasing structure of supersaturated designs. SEAS takes into account interaction terms and provides more informative summaries than traditional design criteria, such as (generalized) resolution and word length patterns, for design evaluations. The new summary consists of three criteria, abbreviated as MAP: (1) the Maximum dependency aliasing (M-)pattern; (2) the Average square aliasing (A-)pattern; and (3) the Pairwise dependency ratio (P-)pattern. We theoretically study the relationships among the three criteria of SEAS and traditional criteria and demonstrate the use of SEAS for evaluating and comparing some examples of supersaturated designs, including those suggested in the literature. We further apply the SEAS to the assignment of columns of a supersaturated design when some important experimental factors are known prior.

C0271: **Criteria for assessing space filling of a design with emphasis on the stratification pattern**
Presenter: **Ulrike Groemping**, Berliner Hochschule fuer Technik, Germany

Designs for computer experiments in quantitative factors should fill the experimental space as well as possible. There are many criteria for the assessment of space-filling properties of a design, and the stratification pattern is a recent addition to this toolbox. It is based on coding for qualitative factors, which renders its calculation computationally demanding. The pattern can be broken down into dimensional contributions, which provide an additional understanding of the design's properties. The logic behind the stratification pattern and its breakdown into dimensions by weight tables are presented. A small simulation study compares the prediction performance of several designs for which the stratification pattern and several further space-filling criteria are considered. Its partly surprising outcomes indicate the need for further research regarding the choice of a design (resp. the choice of a criterion for choosing a design) for actual experimentation.

CO126 Room 44 TUTORIAL: SESSION I Chair: Ivan Savin

C0209: **Topic modelling**
Presenter: **Ivan Savin**, ESCP Business School, Madrid campus, Spain

A theoretical understanding and practical skills for topic modelling will be offered. Using this method, developed on the intersection of machine learning and natural language processing, new ways will be considered about: 1) clustering textual data into meaningful topics, 2) relating this information to other characteristics of the texts and 3) discussing potential ways how to develop a storyline around this type of analysis. Topic modelling can be applied, for example, to responses to open-ended survey questions to elicit preferences, to research articles for large-scale literature reviews, to posts on social networks like Twitter and many more. A few examples of applying the method using R software will be discussed in detail with the participants.

CO078 Room 45 STATISTICAL INFERENCE FOR FUNCTIONAL DATA Chair: Dominik Liebl

C0261: **Factor-augmented functional regression with an application to electricity price curve forecasting**
Presenter: **Sven Otto**, University of Cologne, Germany
Co-authors: Luis Winter

A function-on-function linear regression model is proposed for time-dependent curve data that is consistently estimated by imposing factor structures on the regressors. A novel integral operator based on cross-covariances identifies two components for each functional regressor: a predictive low-dimensional component, along with associated factors that are guaranteed to be correlated with the dependent variable, and an infinite-dimensional component that has no predictive power. In order to consistently estimate the correct number of factors for each regressor, we introduce a functional eigenvalue difference test. The model is applied to forecast electricity price curves in three different energy markets. Its prediction accuracy is found to be comparable to popular machine-learning approaches while providing interpretable insights into the conditional correlation structures of electricity prices.

C0331: **Robust FWER control in neuroimaging using random field theory**
Presenter: **Fabian Telschow**, Humboldt University zu Berlin, Germany
Co-authors: Samuel Davenport, Thomas Nichols, Armin Schwartzman

The Gaussian kinematic formula (GKF) is a computationally efficient tool for performing statistical inference on random fields over large and complex domains. It is also a well-established tool in the analysis of neuroimaging data. The validity of methods based on the GKF, however, has come under scrutiny following a prior seminal work, which utilized error models derived from resting state data collected in the 1000 functional connectomes project. They revealed that while voxelwise inference yields conservative control of the familywise error rate (FWER), cluster-size inference tends to inflate false positive rates. The purpose of this talk is to review the primary factors leading to these findings, notably the unrealistic assumptions regarding "sufficient" smoothness, stationarity, and Gaussianity. Subsequently, a novel method based on the GKF is introduced, which accurately controls the FWER in voxelwise inference. Furthermore, the outcomes of the validation efforts under realistic error models are presented. A big data Eklund style approach is employed, based on resting-state data of 7000 subjects from the UK BioBank.

C0275: **Difference-in-Differences: A Functional data perspective**
Presenter: **Chencheng Fang**, University of Bonn, Germany
Co-authors: Dominik Liebl

Difference-in-Differences (DiD) is typically constructed in a panel data setting, which considers time-series processes in discrete time. We argue, however, that the underlying processes typically can be viewed as (relatively) smooth processes in continuous time, which leads to a functional data perspective. Our theoretical framework takes into account the interpolation errors due to the fact that the underlying functional data are not observed in continuous time, but need to be constructed using natural spline interpolations. We show that the interpolation errors are uniformly negligible under a large n,T asymptotic. It is proved that our functional treatment effects estimator is point-wise asymptotically normal. Moreover, we also adopt a fully functional perspective and show that our functional treatment effects estimator converge to a Gaussian process in the Banach space of continuous functions. The latter result allows us to derive powerful simultaneous confidence bands for the functional treatment effect parameter. One major contribution of our functional perspective on DiD is that we can do functional registration. This provides a completely new possibility for relaxing the critical parallel trends assumption of classic DiD. We show that our registration procedure is consistent, and we derive conditions under which the estimation errors from the registration procedure are asymptotically negligible in the inference about the functional treatment effects.

C0353: **Elastic full Procrustes analysis of plane curves via Hermitian covariance smoothing**
Presenter: **Almond Stoecker**, Ecole polytechnique federale de Lausanne, Switzerland
Co-authors: Sonja Greven, Lisa Steyer, Manuel Pfeuffer

Estimating the mean shape of a collection of curves is a challenging task, particularly when curves are only irregularly/sparsely sampled at discrete points, and so is testing group difference when, due to the sparse sampling, exact distances between curves cannot be computed. An elastic full Procrustes mean of shapes of (oriented) plane curves is newly proposed, which are considered equivalence classes of parameterized curves with respect to the shape invariances translation, rotation and scale, as well as re-parameterization (warping) based on the square-root-velocity (SRV) framework. Identifying the real plane with the complex numbers, a connection to covariance estimation in irregular/sparse functional data analysis is established, and Hermitian covariance smoothing is introduced to employ for (in)elastic full Procrustes mean estimation. Building on this new mean concept and estimator, one- and two-way analysis of variance (ANOVA) is also developed for sparsely sampled curve shape data. The performance of the approach is demonstrated in different realistic simulation settings and is used for an ANOVA of tongue shapes during speech production, taking into account variability in the subject's positioning and size, as well as the elasticity of the tongue.

CC034 Room 001 SEMI- AND NONPARAMETRIC METHODS Chair: Stefan Sperlich

C0197: **A non-parametric approach to detect patterns in binary sequences**
Presenter: **Anushka De**, Indian Statistical Instiute, India

In many circumstances, given an ordered sequence of one or more types of elements/ symbols, the objective is to determine any existence of randomness in the occurrence of one of the elements, say type 1 element. Such a method can be useful in determining the existence of any non-random pattern in the wins or losses of a player in a series of games played. Existing methods of tests based on a total number of runs or tests based on the length of the longest run can be used for testing the null hypothesis of randomness in the entire sequence and not a specific type of element. Additionally, the Runs Test tends to show results that are contradictory to the intuition visualised by the graphs of, say, win proportions over time due to the method used in the computation of runs. A test approach is developed to address this problem by computing the gaps between two consecutive type 1 elements and thereafter following the idea of pattern in occurrence and directional trend (increasing, decreasing or constant), employing the use of exact Binomial test, Kendall's Tau and Siegel-Tukey test for scale problem. Further modifications have been applied in the Siegel Tukey test to adjust for tied ranks and achieve more accurate results. This approach is distribution-free and suitable for small sizes. Also, comparisons with the conventional runs test show the superiority of the proposed approach under the null hypothesis of randomness in the occurrence of type 1 elements.

C0233: **Improved confidence intervals with optimal transportation theory**
Presenter: **Christophe Valvason**, University of Geneva, Switzerland
Co-authors: Stefan Sperlich

Reliable inferential tools for small samples and complex statistics are of high importance in empirical research and official statistics. In this context, we aim to improve the estimation of confidence intervals by using methods developed along the optimal transport theory. When two probability measures satisfy some regularity constraints, the solution to Monge's problem is determined by the composition of the cdf and the quantile function. Clearly, in small samples, the empirical cdf is a step function with a few but large jumps, so the standard transportation map is not necessarily optimal. We propose to compute the optimal transportation plan between a probability distribution of reference that satisfies the regularity constraints and a bootstrap estimate of the probability measure of interest. Then, we construct an exact confidence interval inside the reference distribution and transport this to the problem of interest. Optimal transportation theory provides us with the tools to show the validity of our method. Simulations show that the proposed method gives confidence intervals with coverage closer to the nominal level and at the same time a smaller variance than both, direct and bootstrap confidence interval estimates.

C0494: **L^2-divergence estimator with better finite sample performance**
Presenter: **Sixiao Zhu**, Paris 1 University, France
Co-authors: Alain Celisse

The problem of estimating the L^2 divergence of two continuous probability distributions is studied. A kernel-based estimator is considered where two kernels are employed, providing a finer bias-variance tradeoff for each distribution. Different from the asymptotic regime, where the convergence rate of the whole estimator is dominated by the rougher distribution among the two, in a finite sample regime, the two-kernel framework admits a better oracle estimator than the one-kernel framework. This better oracle is also tractable by a simple model selection procedure, which is shown by providing oracle inequality corresponding to the procedure.

C0466: **Mixed high-dimensional network inference via the Gaussian copula**
Presenter: **Ekaterina Tomilina**, INRAE, France
Co-authors: Gildas Mazo, Florence Jaffrezic

Large-scale heterogeneous data integration for network inference is a key methodological challenge, especially in the context of multi-omic data analysis. A novel procedure is proposed based on the copula theory, which allows the joint analysis of data of various types (continuous, discrete, etc.) The proposed estimation procedure is semi-parametric and, therefore, does not require any explicit assumption concerning the marginal distributions of the data, which offers great flexibility for the analysis of biological data, which may not exactly follow any pre-specified parametric distribution. A theoretical proof is also presented, showing the equivalence between block-wise independence in the copula correlation matrix and the actual data correlation structure. In an extensive simulation study, the proposed estimation procedure is shown, based on a pairwise-pseudo-likelihood approach, was able to accurately estimate the copula correlation matrix, even for a quite large number of variables (several hundred)and a quite small number of replicates (several dozens). The proposed method was also applied to a real ICGC dataset on breast cancer.

CC035 Room 050 CLUSTERING Chair: Maria Brigida Ferraro

C0201: **Non-decimated lifting based outlier detection algorithm**
Presenter: **Nebahat Bozkus**, Giresun University, Turkey

Outlier detection techniques typically generate outlier scores, after which the researcher must establish a threshold to distinguish between inliers and outliers. A novel approach is introduced that assigns empirical probabilities of being outliers to individual objects on a dendrogram using the non-decimated lifting algorithm. The proposed algorithm first removes noise from the hierarchically built tree using the non-decimated lifting transform, then assigns a probability of being an outlier to each object on the tree. Subsequently, the algorithm eliminates objects with high probabilities from the tree and assigns an empirical probability of being a cluster to each node on the tree. This approach is called Non-Decimated Lifting-based Outlier Detection (NDLout). The performance of NDLout is compared with other existing approaches in the literature using real-world and synthetic datasets.

C0397: **Divisive hierarchical clustering of variables identified by singular vectors**
Presenter: **Jan Bauer**, Vrije Universiteit Amsterdam, Netherlands

A novel method is presented for divisive hierarchical variable clustering. A cluster is a group of elements that exhibit higher similarity among themselves than to elements outside this cluster. The correlation coefficient serves as a natural measure to assess the similarity of variables. This means that in a correlation matrix, a cluster is represented by a block of variables with greater internal than external correlation. The approach provides a nonparametric solution to identify such block structures in the correlation matrix using singular vectors of the underlying data matrix. When divisively clustering p variables, there are 2^{p-1} possible splits. Using the singular vectors for cluster identification, these numbers can be effectively reduced to at most $p(p-1)$, thereby making it computationally efficient. The methodology is elaborated, and the incorporation of dissimilarity measures and linkage functions is outlined to assess distances between clusters. Additionally, it is demonstrated that these distances are ultrametric, ensuring that the resulting hierarchical cluster structure can be uniquely represented by a dendrogram, with the heights of the dendrogram being interpretable. To validate the efficiency of the method, simulation studies are performed, and real-world data on personality traits and cognitive abilities is analyzed.

C0404: **Modified silhouette score for evaluating cluster solutions**

Presenter: **Czarinne Antoinette Antonio**, University of the Philippines Diliman, Philippines

Co-authors: Joseph Ryan Lansangan

The assessment of the quality of a clustering solution and the proper identification of the number of clusters to be used is a crucial step in doing cluster analysis. A class of silhouette-based indices, as a modification to the widely used silhouette index, is developed to measure cluster validity. The performance of the proposed indices is demonstrated via a simulation study and through the application to actual data sets. The results revealed that the use of the second and third nearest cluster in the computation instead of just the nearest neighboring cluster relative to observation was advantageous in identifying the number of natural clusters as a viable choice in the cluster analysis. Each of the proposed indices was useful in the presence of noisy data and not well-separated clusters. Further, dimension reduction techniques employed in the calculation of the distance measures provided an added benefit when dealing with high-dimensional data.

C0437: **A new approach to estimate semi-parametric Gaussian mixtures of regressions with varying mixing proportions**

Presenter: **Sphiwe Skhosana**, University of Pretoria, South Africa

Co-authors: Sollie Millard, Frans Kanfer

The semi-parametric Gaussian mixture of regressions with varying proportions (SPGMRVPs) model is a flexible version of a Gaussian mixture of linear regressions (GMLRs) model. The model assumes that the mixing probabilities are non-parametric functions of the covariate(s). Traditional methods of estimation are not guaranteed to produce reliable estimates of the model. A local-likelihood approach for estimating the non-parametric functions requires that we maximize a set of local-likelihood functions. Using the Expectation-Maximization (EM) algorithm to separately maximize each local-likelihood function may lead to label switching. This is because the responsibilities calculated at each local E-step might not be aligned. The consequence of this label-switching is wiggly and non-smooth non-parametric functions as a result of incorrect component identification. We propose a novel approach to address label-switching and obtain improved estimates. We propose a model-based approach to address the label-switching problem. We reformulate the SPGMRVPs model as a mixture of local GMLRs. Estimating the mixture of GMLRs is equivalent to simultaneously maximizing the local-likelihood functions. Next, we propose one-step backfitting estimates of the parametric and non-parametric terms. The effectiveness and practical utility of the proposed approach is demonstrated using Monte Carlo simulations and environmental data analysis.

CC139 Room 051 NEURAL NETWORKS Chair: Florian Brueck

C0187: **Generative neural networks for characteristic functions**

Presenter: **Florian Brueck**, University of Geneva, Switzerland

A simulation algorithm is provided to simulate from a (multivariate) characteristic function, which is only accessible in a blackbox format. We construct a generative neural network, whose loss function exploits a specific representation of the Maximum-Mean-Discrepancy metric to directly incorporate the targeted characteristic function. The construction is universal in the sense that it is independent of the dimension and that it does not require any assumptions on the given characteristic function. Furthermore, finite sample guarantees on the approximation quality in terms of the Maximum-Mean Discrepancy metric are derived. The method is illustrated in a short simulation study

C0449: **Learning of deep convolutional network image classifiers via stochastic gradient descent and over-parametrization**

Presenter: **Alisha Saenger**, Technische Universitat Darmstadt, Germany

Co-authors: Michael Kohler, Adam Krzyzak

Image classification from independent and identically distributed random variables is considered. Image classifiers are defined based on a linear combination of deep convolutional networks with a max-pooling layer. All the weights are learned by stochastic gradient descent. A general result is presented, which shows that the image classifiers are able to approximate the best possible deep convolutional network. In case the a posteriori probability satisfies a suitable hierarchical composition model, it is shown that the corresponding deep convolutional neural network image classifier achieves a rate of convergence independent of the dimension of the images.

C0434: **Nonparametric estimation of conditional class probabilities using deep neural networks**

Presenter: **Atsutomo Yara**, Graduate School of Engineering Science, Osaka University, Japan

Co-authors: Yoshikazu Terada

Consider the nonparametric logistic regression problem. In the logistic regression, the maximum likelihood estimator is usually considered, and the excess risk is the expectation of the Kullback-Leibler (KL) divergence between the true and estimated conditional class probabilities. However, in the nonparametric logistic regression, the KL divergence could diverge easily, and thus, the convergence of the excess risk is difficult to prove or does not hold. Several existing studies show the convergence of the KL divergence under strong assumptions. In most cases, the goal is to estimate the true conditional class probabilities. Thus, instead of analyzing the excess risk itself, it suffices to show the consistency of the maximum likelihood estimator in some suitable metric. Using a simple unified approach for analyzing the nonparametric maximum likelihood estimator (NPMLE), the convergence rates of the NPMLE are directly derived in the Hellinger distance under mild assumptions. Although the results are similar to the results in some existing studies, simple and more direct proofs are provided for these results. As an important application, the convergence rates of the NPMLE are derived with deep neural networks, and the derived rate is shown to achieve the minimax optimal rate nearly.

Friday 30.08.2024 11:00 - 12:00 Parallel Session N – COMPSTAT2024

CO106 Room 052 CAUSAL MACHINE LEARNING Chair: Martin Spindler

C0183: **Comprehensive causal machine learning**
Presenter: **Michael Lechner**, University St Gallen, Switzerland
Co-authors: Jana Mareckova

Uncovering the heterogeneity of causal effects at various levels of granularity provides substantial value to decision-makers. Comprehensive approaches to causal effect estimation allow the use of a single causal machine learning approach for the estimation and inference of causal mean effects for all levels of granularity. Focussing on selection-on-observables, the theoretical asymptotic guarantees for one such approach, the modified causal forest (mcf), are provided. The asymptotic and finite sample properties of the mcf to the generalized random forest (rf) and double machine learning (DML) are also compared. The findings indicate that dml-based methods excel for average treatment effects at the population level (ATE) and group level (GATE) with few groups. However, for finer causal heterogeneity, explicitly outcome-centred forest-based approaches are superior. The mcf has three additional benefits: (i) It is the most robust estimator in cases when dml-based approaches underperform because of substantial selectivity; (ii) it is the best estimator for GATEs when the number of groups gets larger; and (iii), it is the only estimator that is internally consistent, in the sense that low-dimen-sional causal ATEs and GATEs are obtained as aggregates of finer-grained causal parameters.

C0205: **Testing identification in mediation and dynamic treatment models**
Presenter: **Kevin Kloiber**, LMU Munich, Germany
Co-authors: Martin Huber, Lukas Laffers

A test is proposed for the identification of causal effects in mediation and dynamic treatment models based on two sets of observed variables, namely covariates to be controlled for and suspected instruments, building on a previous test for single treatment models. We consider models with a sequential assignment of a treatment and a mediator to assess the direct treatment effect (net of the mediator), the indirect treatment effect (via the mediator), or the joint effect of both treatment and mediator. We establish testable conditions for identifying such effects in observational data. These conditions jointly imply (1) the exogeneity of the treatment and the mediator conditional on covariates and (2) the validity of distinct instruments for the treatment and the mediator, meaning that the instruments do not directly affect the outcome (other than through the treatment or mediator) and are unconfounded given the covariates. The framework extends to post-treatment sample selection or attrition problems when replacing the mediator with a selection indicator for observing the outcome, enabling joint testing of the selectivity of treatment and attrition. We propose a machine learning-based test to control for covariates in a data-driven manner and analyze its finite sample performance in a simulation study. We apply our method to Slovak labor market data and find that our testable implications are not rejected for a typical sequence of training programs

C0415: **Multivariate binary extension for W&A-learner**
Presenter: **Shintaro Yuki**, Doshisha University, Japan
Co-authors: Kensuke Tanioka, Hiroshi Yadohisa

Randomized controlled trials and observational studies are conducted to test the efficacy of a treatment, and the focus is on two-arm comparisons. The results of a treatment may not demonstrate efficacy in a population that meets the eligibility criteria. In such cases, it is desirable to efficiently identify populations with characteristics that make the treatment effective so-called subgroups. They can be identified by estimating the treatment effect. A prior study introduced the W-learner and A-learner as approaches for modeling the interaction between treatment modalities and covariates using propensity score weighting aimed at estimating the effect of treatments on a single outcome. Subsequently, a recent study extended the W-learner to handle multivariate outcomes. However, while this method accounts for the correlation structure among multiple continuous outcomes, it fails to address the correlation structure among multiple binary outcomes, and no multivariate extension has been applied to the A-learner. A novel approach that enables both the W-learner and A-learner to account for the correlation structure among multiple binary outcomes is proposed.

CO103 Room 43 HITEC: REGRESSION FUNCTIONS Chair: Matus Maciak

C0399: **Changepoints in a nonlinear expectile model: Theoretical and computational issues**
Presenter: **Matus Maciak**, Charles University, Czech Republic

An online instability detection test proposed to detect changepoints in a nonlinear expectile model is discussed. Conditional expectiles, well-known in econometrics for being the only coherent and elicitable risk measure, introduce some portion of robustness in the underlying model, and the proposed statistical test is proved to be consistent while the distribution of the test statistic under the null hypothesis does not depend on the functional form of the underlying model. Resampling techniques are used to obtain the final test decision, and, therefore, relatively easy and straightforward practical application is guaranteed. Important theoretical details are discussed, and finite sample empirical properties and real data illustrations are presented.

C0438: **Inference on derivatives of high-dimensional regression function with deep neural networks (NN)**
Presenter: **Weining Wang**, University of Groningen, Netherlands

The purpose is to study the estimation of the partial derivatives of non-parametric regression functions with many predictors and a subsequent significance test for the said derivatives. The derivative estimator is the derivative of the convolution of a regression function estimator and a smoothing kernel, where the regression function estimator is a deep neural network whose structure could scale up as the sample size grows. It is demonstrated that in the context of modeling with deep neural networks, derivative estimation is quite different from estimating the regression function itself, and hence the smoothing operation becomes beneficial and even necessary. The subsequent significance test, where the null hypothesis is that a partial derivative is zero, is based on the moment-generating function of the aforementioned derivative estimator. This test finds applications in model specification and variable screening for high-dimensional data. To render the estimator and test effective when facing predictors with high or even diverging dimensions, it is assumed that first, the observed high-dimensional predictors can effectively serve as the proxies for certain latent, lower-dimensional factors and that second, only the latent factors and a subset of the coordinates of the observed high-dimensional predictors drive the regression function.

C0505: **Shrinkage estimation for multivariate linear regression**
Presenter: **Vali Asimit**, City University London, United Kingdom

Shrinkage estimation is a widely popular estimation procedure that is triggered by Stein's paradox, which had puzzled the statistical community for some time. The intuition behind is that different sources of information could be combined to "better" deal with a multidimensional estimation problem rather than separately estimate the individual estimation problems. We consider four shrinkage estimators to make "better" predictions for a multivariate linear regression model. Our simulation results and real data analyses show that our four shrinkage estimators are no worse than the classical Ordinary Least Square estimator, though at least one shrinkage estimator significantly improves the performance of the classical estimator. For example, the generalised linear model estimation massively benefits from our new estimators.

CO127 Room 44 TUTORIAL: SESSION II Chair: Ivan Savin

C0152: **Topic modelling**

Presenter: **Ivan Savin**, ESCP Business School, Madrid campus, Spain

A theoretical understanding and practical skills for topic modelling will be offered. Using this method, developed on the intersection of machine learning and natural language processing, new ways will be considered about: 1) clustering textual data into meaningful topics, 2) relating this information to other characteristics of the texts and 3) discussing potential ways how to develop a storyline around this type of analysis. Topic modelling can be applied, for example, to responses to open-ended survey questions to elicit preferences, to research articles for large-scale literature reviews, to posts on social networks like Twitter and many more. A few examples of applying the method using R software will be discussed in detail with the participants.

CO120 Room 45 STOCHASTIC SIMULATION IN COMPUTATIONAL STATISTICS Chair: David Fernando Munoz

C0305: **Simulation models for planning the electoral results program for the 2024 federal elections in Mexico**

Presenter: **David Fernando Munoz**, Instituto Tecnologico Autonomo de Mexico, Mexico

Since the 2018 federal elections, the Technical Unit of Computer Services of the National Electoral Institute of Mexico has used simulation models as planning tools for the execution, on the day of the elections, of the Preliminary Electoral Results Programs (PREP). For the 2024 federal elections, a simulation model was developed using the commercial software Simio. This model aims to predict the percentage of scrutiny forms published per count update (after the closure of polling booths) under different capacities of the resource pools that execute the different activities and has the capacity to simulate the exchange of resources between the different pools, according to the processing needs at different moments in the development of the processes. Due to the large number of entities and resources involved in the PREP technical operational process (PTO), the model developed in Simio requires very long running times, so it was necessary to replicate this model using subroutines specifically developed in C++ for the PREP PTO. This second model runs in significantly shorter times than the Simio model, and allowed us to obtain the results required for planning of the PREP for the 2024 federal elections. Both the development and performance of the simulation models are reported. as planning tools, and details are provided on the steps followed in building the model, such as input analysis, model verification and animation, and output analysis

C0317: **Simulation and assessment of Subway Line 7 in Mexico City public transportation**

Presenter: **Elias Arias Nava**, Instituto Tecnologico Autonomo de Mexico (ITAM), Mexico

Mexico City is a massive city with millions of people using public transportation every day. According to the latest census, more than 9 million people live in the metropolitan area. Public transportation is one of the government's most pressing commitments. Starting with the development and construction of the first subway line in 1969, Mexico has accumulated 226 kilometers of subway railroads distributed among 12 lines. The goal is to analyze and assess the problems of the subway: logistics, efficiency, capacity, waiting times, and overall capability to attend to all customers. Using system simulation methodology: problem formulation, setting of objectives, model conceptualization and data collection, model translation, verification and validation, experimental design, production runs and analysis, and documentation; statistically significant information was collected and analyzed in terms of the performance of the subway system. The database used as input parameters included information from one year, that is, the ride information of approximately 800 million passengers. Scheduling, distribution, and user decision-making when using the system are part of the solution and recommendations based on the quantitative results from the simulation output parameters.

C0382: **Identification of cutting model parameters for a flat milling process**

Presenter: **Thomas Martin Rudolf**, Instituto Tecnologico Autonomo de Mexico - ITAM, Mexico

Process monitoring for milling operations is widely used to ensure product quality and optimize costs. Detecting a worn tool on time can prevent tool breakage and quality issues. Therefore, the observation of wear status is crucial in process monitoring. A typical approach to tool wear detection is the monitoring of required energy and resulting cutting forces. The forces are based on the removed volume and the material characteristics. In the first part, the author explains different cutting force models, their corresponding parameters, and their impact on modelled forces. The force is typically defined by $Fc = kc1.1bh^{(1-m_c)}$, with parameters specific cutting force $k_{c1.1}$ and increasing value of the specific cutting force $(1-mc)$, b and h are the geometric values of the removed material, width and height, respectively. The significance of the parameters $k_{c1.1}$ and $(1-m_c)$ are the topic of discussion. Although there are known values for specific materials, each working batch has slightly different values, which results in changing force values for the same machining process. The presented approach uses a Bayesian method to detect and identify the current values based on former knowledge. A prior distribution for both parameters is defined, and their selection is explained. Then, new data is acquired during the first cuts, and the resulting distribution is calculated.

CC015 Room 001 COMPUTATIONAL AND FINANCIAL ECONOMETRICS Chair: Francesco Audrino

C0428: **HARd to beat: The overlooked impact of rolling windows in the era of machine learning**

Presenter: **Jonathan Chassot**, University of St.Gallen, Switzerland

Co-authors: Francesco Audrino

The predictive abilities of the heterogeneous autoregressive (HAR) model compared to machine learning (ML) techniques are investigated across an unprecedented dataset of 1'445 stocks. The analysis focuses on the role of fitting schemes, particularly the training window and re-estimation frequency, in determining the HAR model's performance. Despite extensive hyperparameter tuning, ML models fail to surpass the linear benchmark set by HAR when utilizing a refined fitting approach for the latter. Moreover, the simplicity of HAR allows for an interpretable model with drastically lower computational costs. Performance is assessed using QLIKE, MSE, and realized utility metrics, finding that HAR consistently outperforms its ML counterparts when both rely solely on realized volatility and VIX as predictors. The results underscore the importance of a correctly specified fitting scheme. They suggest properly fitted HAR models provide superior forecasting accuracy, establishing robust guidelines for their practical application and use as a benchmark. The efficacy of the HAR model is not only reaffirmed but also a critical perspective on the practical limitations of ML approaches is provided in realized volatility forecasting.

C0328: **A multi-factor model for pricing commodities when volatility, interest rate and convenience yield are stochastic**

Presenter: **Christian Tezza**, University of Bologna, Italy

Co-authors: Luca Vincenzo Ballestra

A novel affine model is introduced for commodity pricing that builds on previous well-known approaches, incorporating four stochastic uncertainty factors. The proposed model allows for both the volatility and long-run mean of commodity prices, as well as the instantaneous convenience yield and interest rate, to be stochastic. The Kalman filter is utilized in conjunction with quasi-maximum likelihood estimation to estimate the model parameters. An empirical analysis that focuses on different commodity futures is conducted to evaluate the performance of the novel specification. The model's in-sample and out-of-sample results are compared to existing approaches. The approach can help enhance risk management strategies and improve decision-making in commodity markets.

C0349: **Dense-to-sparse neural network modelling for financial statement data using feature importance attributions**

Presenter: **Lars Fluri**, University of Basel, Switzerland

A new approach is proposed to feature selection and sparse modelling in the context of financial data analysis to predict free cash flow. Utilising deep learning important features (DeepLIFT), a process for iterative elimination of input features is introduced, reducing the model complexity and enhancing the robustness through the elimination of less significant input nodes. Furthermore, a method for the regrowth of nodes using the gradient magnitude of previously eliminated features is used. Drawing on a dataset of 874 firms from the DACH region over a decade, the model is used to identify forward-looking predictors of free cash flow. Additionally, it evaluates both computational aspects and performance metrics (including in-sample and out-of-sample performance) to measure improvements from the original dense model to the optimized sparse model. The contribution is to the evolving field of machine learning applications in finance, proposing an alternative framework for feature selection and model optimization.

CC031 Room 050 STATISTICAL MODELLING Chair: Tatyana Krivobokova

C0411: **Copula estimation with flow copula models**

Presenter: **Bolin Liu**, Ludwigshafen University of Business and Society, Germany

Co-authors: Oliver Grothe, Maximilian Coblenz

Flow copulas are introduced as a new copula class based on the change of variables formula. Theoretical properties such as the universal expressive power, i.e. any well-behaved absolutely continuous copula can be modeled by a flow copula, are shown. Furthermore, constructions of flow copula models based on the normalizing flow technique and its training procedures are presented. For this, a customized model structure is developed that guarantees copula properties such as uniform margins, which enables the estimation of a flow copula model from data. The learned flow copula model can then be used to estimate the copula density or to generate synthetic data. In simulation studies, it is shown that the presented flow copula models not only can represent various dependence properties such as tail dependence and asymmetry but also provide comparable results to other non-parametric copula estimation methods. Furthermore, the proposed model is applied to real data.

C0413: **A generalized voting game for categorical network choices**

Presenter: **Yueh Lin**, IESEG School of Management, France

Co-authors: Stefano Nasini, Martine Labbe

A game theoretical framework is presented for data classification based on the interplay of pairwise influences in multivariate choices. This consists of a voting game wherein individuals, connected through a weighted network, select features from a finite list. A voting rule captures the positive or negative influence of an individual's neighbours, categorized as attractive (friend-like relationships) or repulsive (enemy-like relationships). Payoffs are assigned based on the total number of matching choices from an individual's neighbours. It is shown that the approach constitutes a natural generalization of the K-nearest neighbors method, establishing the proposed game as a theoretical framework for data classification. Computationally, a mixed-integer linear programming formulation is constructed to approach the Nash equilibria of the game, facilitating their applicability to real-world data. The results provide conditions for the existence of Nash equilibria and for the NP-completeness of its characterization. On the empirical side, the proposed approach is used to impute missing data and highlight its competitive advantage over the K-nearest neighbors approach.

C0439: **The four-parameter exponentiated Weibull exponential distribution: Theoretical properties and practical implications**

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

Co-authors: Patricia Antunes, Dario Ferreira

The four-parameter exponentiated Weibull exponential distribution is introduced, and its different statistical properties are studied. The proposed distribution has several desirable properties and covers some existing distributions. Some important statistical properties of this distribution are obtained, like moments, cumulants, and estimators, which are fundamental to understanding the distribution's characteristics. The estimation of the model parameters is also considered.

CC066 Room 051 SURVIVAL ANALYSIS Chair: Shu-Kay Angus Ng

C0388: **Inequalities and bounds for order statistics**

Presenter: **Paulo Oliveira**, CWT - Turistrader - Sociedade de Desenvolvimento Turistico, Lda, Portugal

The lifetime of k-out-n systems is described by order statistics. Stochastic ordering of order statistics is a study exploring shape properties of the underlying base distribution, expressed via stochastic comparison with some reference distributions, that define a hierarchy of classes, characterizing a relaxing chain of shape restrictions on the base distribution. One of the classes is the popular increasing hazard rate family, which may be defined by stochastic comparison with the exponential distribution. The method and results are extended to a broader class comparing distributions with respect to the log-logistic distributions, which also show nice lack-of-memory properties. The inclusion in the proposed classes is also tested. Finally, as a consequence of inclusion in appropriate families of distributions, explicit conditions are still derived for comparing moments of order statistics, and, with an application of the Jenssen inequality, some bounds for suitable exceedance probabilities.

C0432: **Assessing significance of covariates in mixture cure models using distance correlation**

Presenter: **Blanca Monroy-Castillo**, Universidade da Coruna, Spain

Co-authors: Ingrid Van Keilegom, M Amalia Jacome, Ricardo Cao

One of the challenges in cure models is to test whether a covariate influences the cure rate. Distance correlation is a novel class of multivariate dependence coefficients that offers advantages over classical correlation coefficients: it is applicable to random vectors of arbitrary dimensions, not necessarily equal, and it is zero if and only if the vectors are independent. Distance correlation has been applied in a standard survival model without a cure based on the distance covariance between covariates and survival times. However, to the best of our knowledge, distance correlation has not yet been applied in the presence of a cure fraction. One challenge in dealing with cure survival data is that the cure indicator is only partially observed due to censoring. A method to study the effect of a covariate on the probability of cure using distance correlation is proposed, which overcomes the challenge of handling the missingness of the cure indicator.

C0501: **Credit scoring using varying coefficients survival models**

Presenter: **Viani Djeundje**, University of Edinburgh, United Kingdom

Credit scoring models constitute a major instrument used by financial institutions to evaluate the risk associated with a loan. At its core, a credit scoring model involves predicting the probability that an account will default over a future time period based on a number of observed variables or attributes that characterize account holders or applicants. Traditional scoring methods were based essentially on the attributes of the applicants measured at the time of application. Yet, many characteristics of the applicants change with time. Survival analysis techniques provide an attractive platform to address the limitations of traditional methods. However, most applications of survival models encountered in the literature assume that the impact of each risk factor on the probability of default remains constant over the business cycle. The purpose is to investigate the validity of such an assumption in the context of retail banking using a large portfolio of credit card loans from a major UK bank. Specifically, a class of flexible models is considered in which the relative impacts of the risk factors are free to vary. A parametric formulation and a spline specification are then proposed to capture the dynamic patterns of the impacts of these risk factors over time. Finally, the varying coefficient approach is shown to outperform a standard model in terms of overall model quality and prediction accuracy.

Friday 30.08.2024 12:10 - 13:00 Parallel Session O – COMPSTAT2024

CO128 Room 44 TUTORIAL: SESSION III Chair: Ivan Savin

C0155: **Topic modelling**

Presenter: **Ivan Savin**, ESCP Business School, Madrid campus, Spain

A theoretical understanding and practical skills for topic modelling will be offered. Using this method, developed on the intersection of machine learning and natural language processing, new ways will be considered about: 1) clustering textual data into meaningful topics, 2) relating this information to other characteristics of the texts and 3) discussing potential ways how to develop a storyline around this type of analysis. Topic modelling can be applied, for example, to responses to open-ended survey questions to elicit preferences, to research articles for large-scale literature reviews, to posts on social networks like Twitter and many more. A few examples of applying the method using R software will be discussed in detail with the participants.

Friday 30.08.2024 14:00 - 16:30 Parallel Session P – COMPSTAT2024

CO129 Room 44 TUTORIAL: SESSION IV Chair: Ivan Savin

C0157: **Topic modelling**

Presenter: **Ivan Savin**, ESCP Business School, Madrid campus, Spain

A theoretical understanding and practical skills for topic modelling will be offered. Using this method, developed on the intersection of machine learning and natural language processing, new ways will be considered about: 1) clustering textual data into meaningful topics, 2) relating this information to other characteristics of the texts and 3) discussing potential ways how to develop a storyline around this type of analysis. Topic modelling can be applied, for example, to responses to open-ended survey questions to elicit preferences, to research articles for large-scale literature reviews, to posts on social networks like Twitter and many more. A few examples of applying the method using R software will be discussed in detail with the participants.

Authors Index