PROGRAMME AND ABSTRACTS

HiTEc meeting & Workshop on

Complex data in Econometrics and Statistics (HiTEc & CoDES 2024)

`https://www.cmstatistics.org/hiteccodes2024`

Cyprus University of Technology, Limassol, Cyprus
23-24 March 2024

HiTEc meeting &
Workshop on Complex data in
Econometrics and Statistics

Contents

Saturday 23.03.2024 09:00 - 09:50 Room: Amphitheater 1 Chair: Erricos Kontoghiorghes Keynote talk 1

Combinatorial algorithms for variable selection in regression

Speaker: **Cristian Gatu, University of Iasi, Romania** M Hofmann, Marios Demosthenous, Erricos Kontoghiorghes

Computational strategies for computing the best-subset regression models are proposed. The algorithms are based on a regression tree structure that generates all possible subset models. An efficient branch-and-bound algorithm that finds the best submodels without generating the entire tree is described. Approximate algorithms that improve the computational performance are investigated. Further, this strategies are adapted to solve the problem of regression subset selection under the condition of non-negative coefficients. The solution is based on an alternative approach to quadratic programming that derives the non-negative least squares by solving the normal equations for a number of unrestricted least squares subproblems. The R package "lmSubsets" for regression subset selection is introduced and described. The package aims to provide a versatile tool for subset regression. Finally, the case of high-dimensional data where the number of variables exceeds the number of observations is considered. Within this context, a novel combinatorial solution is proposed. It generates a high-dimensional regression tree to select the optimal model of size up to k variables, where k is smaller than the available observation in the data. It avoids evaluating the same model more than once and utilizes previous computations for evaluating subsequent combinations of variables, thus reducing the computational cost. Experimental results are presented and analyzed.

Sunday 24.03.2024 17:35 - 18:25 Room: Amphitheater 1 Chair: Ana Colubi Keynote talk 2

Fuzzy clustering: From numerical to complex data

Speaker: **Maria Brigida Ferraro, Sapienza University of Rome, Italy**

The fuzzy approach to clustering arises to cope with situations where objects do not have a clear assignment. Unlike the hard/standard approach, where each object can only belong to exactly one cluster, in a fuzzy setting, the assignment is soft; that is, each object is assigned to all clusters with certain membership degrees varying in the unit interval. The best-known fuzzy clustering algorithm is the fuzzy k-means (FkM) or fuzzy c-means. It is a generalization of the classical k-means method. Starting from the FkM algorithm, and in more than 40 years, several variants have been proposed. The peculiarity of such different proposals depends on the type of data to deal with and on the cluster shape. The aim is to show fuzzy clustering alternatives to manage different kinds of data, ranging from numerical, categorical or mixed data to more complex data structures, such as interval-valued, fuzzy-valued or network data, together with some robust methods. Furthermore, the case of two-mode clustering is illustrated in a fuzzy setting.

Saturday 23.03.2024 10:20 - 12:25 Parallel Session B – HiTECCoDES2024

HI028 Room Amphitheater 2 ELECTRICITY MARKETS AND APPLIED MACHINE LEARNING Chair: Christina Erlwein-Sayer

H0163: **Day-ahead probability forecasting for redispatch 2.0 measures**

Presenter: **Alla Petukhina**, HTW Berlin, Germany

Co-authors: Christina Erlwein-Sayer, Mai Phan, Maria Basangova, Alexandra Conda, Vlad Bolovaneanu, Awdesch Melzer

The purpose is to advance a data-driven, day-ahead forecasting model for assessing the probability, direction, and scale of electrical congestions within the German complex power grid. Utilizing state-of-the-art machine learning algorithms, the model is specifically designed to operate on an hourly basis, thereby offering timely insights for grid management. The analysis uncovers compelling evidence that key exogenous variables, such as real-time meteorological conditions, electricity supply-demand indicators, and Brent oil price fluctuations, can be harnessed to make highly reliable predictions concerning grid congestion events. The model has the potential to serve as a useful resource for transmission system operators (TSOs) and policymakers interested in grid management and cost mitigation efforts.

H0172: **Regime shifts in LSTM models for spot prices**

Presenter: **Christina Erlwein-Sayer**, University of Applied Sciences HTW Berlin, Germany

Co-authors: Tilman Sayer, Florian Schirra, Stefanie Grimm

Electricity spot prices show volatile periods and frequently occurring jumps over time. A prediction of day-ahead spot prices relies on suitable modelling paradigms to capture these changing dynamics. A regime-switching HMM is developed, which drives a decision-making process consisting of long-term memory models (LSTM) for specific time periods. Market regimes are adaptively filtered out from the data set and utilized to split the spot price series. This leads to the n-state HMM-LSTM, which is trained on split regime-specific daily prices. Weighted LSTM estimates lead to day-ahead predictions. Combining LSTM with filtered Markov chain probabilities increases the interpretability of predictions. Each activated LSTM is dependent on the filtered state of the underlying market. The model is applied to an extensive data set of German spot prices.

H0178: **Understanding the drivers of electricity prices in the day-ahead market: A factor analysis approach**

Presenter: **Eleftheria Paschalidou**, Aristotle University of Thessaloniki, Greece

Co-authors: Nikolaos Thomaidis

The purpose is to investigate the dynamics between day-ahead electricity prices and their underlying fundamental drivers in the Spanish day-ahead market. Blending structural dynamic factor models (SDFM) with fractionally integrated vector autoregressive (FIVAR) techniques, it unravels complex relationships between hourly prices and most of their key determinants (average surface temperature, load, renewable energy injection, conventional power natural gas price and the CO2 emission rights market value). At the core of the analysis lies the multi-level factor modelling device that explains the covariation of observables based on their exposure to common orthogonal (lagged uncorrelated) components. Subsequently, the FIVAR model captures long memory effects in the dynamic interactions between electricity prices and fundamental drivers. Preliminary results underscore the effectiveness of this integrated approach in capturing nuanced dynamics within the Spanish electricity market. Notably, the analysis reveals that fundamental shocks are absorbed slowly (at a hyperbolic rate), a finding which is consistent with the long-range dependency behavior typically observed in energy indices. Overall, results support the hypothesis that variations in wholesale day-ahead electricity prices are primarily rational, i.e. they can be effectively explained by shifts in fundamental drivers.

H0162: **A hybrid machine learning framework for tax revenues monitoring**

Presenter: **Daria Scacciatelli**, Sogei, Italy

Co-authors: Eugenio Cangiano, Francesco De Napoli

Monitoring tax revenues at least monthly is extremely important for assessing the convergence of public finance figures with annual objectives. This is especially relevant due to the potential revisions in annual budget forecasts by the Italian Ministry of Economy and Finance, influenced by changes in fiscal policy measures or updates in macroeconomic scenarios. To assess the impact of such revisions, a higher-frequency model is proposed, incorporating additional information gathered throughout the year. The proposed hybrid machine learning framework, named HGB, rooted in the gradient boosting algorithm, is designed to generate short-term forecasts of tax revenues. This framework integrates feature selection methods, auto-regressive models, and Machine Learning regression algorithms. Data from diverse sources are gathered and directed to a centralized data hub, where the Boruta algorithm identifies relevant information. The SARIMA model predicts future values for selected variables, and the XGBoost model uses these predictions to derive tax revenue forecasts. In the experimental results, the analysis focuses on excise duty on mineral oil, representing one of the indirect taxes. The HGB framework exhibited high predictive accuracy, outperforming traditional autoregressive models used as benchmarks. The evaluation was performed using the k-fold cross-validation method.

H0171: **Application of machine learning methods to forecast cryptocurrencies volatility**

Presenter: **Witold Orzeszko**, Nicolaus Copernicus University in Torun, Poland

Co-authors: Piotr Fiszeder, Grzegorz Dudek, Pawel Kobus

A comprehensive study of statistical and machine learning methods is presented for predicting the volatility of the following four cryptocurrencies: Bitcoin, Ethereum, Litecoin, and Monero. Several methods, i.e., HAR, ARFIMA, GARCH, LASSO, ridge regression, SVR, MLP, fuzzy neighbourhood model, random forest, and LSTM, are compared in terms of their forecasting accuracy. The realized variance calculated from intraday returns is used as the input variable for the models. The experimental results demonstrate that there is no single best method for forecasting the volatility of each cryptocurrency, and different models may perform better depending on the specific cryptocurrency, choice of the error metric and forecast horizon. Furthermore, it is shown that simple linear models such as HAR and ridge regression do not perform worse than more complex models like LSTM and random forest. The research provides a useful reference point for the development of more complex models and suggests the potential benefits of incorporating additional input variables.

Saturday 23.03.2024 14:00 - 15:40 Parallel Session C – HiTECCoDES2024

HO017 Room Amphitheater 1 ADVANCES ON HIGH-DIMENSIONAL AND COMPLEX DATA Chair: Eugen Pircalabelu

H0177: **Asymptotic normality and bias correction in high-dimensional statistical inference with regularized estimators**
Presenter: **Jing Zhou**, University of East Anglia, United Kingdom

Bias correction stands as an important technique for high-dimensional statistical inference using regularized estimators. The idea of this technique is to correct the bias caused by the regularizer and demonstrate an asymptotic normality of a complete sparse parameter vector. This line of research is advantageous in that it considers selection uncertainty and allows for variability of the nonnull components of the parameter vector. This is especially attractive because the asymptotic oracle properties of the regularized estimators are unlikely to hold in finite samples. Obtaining the asymptotic normality of the biased corrected regularized estimators provides flexibility in both estimation and variable selection. The potential of hypothesis testing is showcased using the bias correction technique for the regularized M-estimators.

H0188: **Turnstile ℓ_p leverage score sampling with applications**
Presenter: **Alexander Munteanu**, TU Dortmund, Germany
Co-authors: Simon Omlor

The turnstile data stream model offers the most flexible framework where data can be manipulated dynamically, i.e., rows, columns, and even single entries of an input matrix can be added, deleted, or updated multiple times in a data stream. A novel algorithm is developed for sampling rows a_i of a matrix $A \in \mathbb{R}^{n\times d}$, proportional to their ℓ_p norm when A is presented in a turnstile data stream. The algorithm not only returns the set of sampled row indexes, but it also returns slightly perturbed rows $\tilde{a}_I \approx a_i$, and approximates their sampling probabilities up to ε relative error. When combined with preconditioning techniques, the algorithm extends to ℓ_p leverage score sampling over turnstile data streams. With these properties in place, it allows the simulation of subsampling constructions of coresets for important regression problems to operate over turnstile data streams with very little overhead compared to their respective off-line subsampling algorithms. For logistic regression, this framework yields the first algorithm that achieves a $(1+\varepsilon)$ approximation and works in a turnstile data stream using polynomial sketch/subsample size, improving over $O(1)$ approximations or $\exp(1/\varepsilon)$ sketch size of previous work.

H0190: **(Almost) real time outlier detection via principal least squares support vector machines**
Presenter: **Andreas Artemiou**, University of Limassol, Cyprus

The aim is to propose an influence measure for outlier detection using principal least squares support vector machines (PLSSVM). A number of papers have discussed the influence measure of sufficient dimension reduction (SDR) methodology, but they focus on inverse moment-based (SDR) methods. The influence measure for an SVM-based SDR method is developed for the first time. Also, given that the PLSSVM algorithm was originally proposed in the online dimension reduction framework, it is demonstrated that this influence measure can be applied in an online outlier detection framework.

H0205: **Direction identification and minimax estimation by generalized eigenvalue problem in high dimensional sparse regression**
Presenter: **Mathieu Sauvenier**, Universite Catholique de Louvain, Belgium
Co-authors: Sebastien Van Bellegem

In high-dimensional sparse linear regression, the selection and the estimation of the parameters are studied based on an l_0-constraint on the direction of the vector of parameters. A general result for the direction of the vector of parameters is first established, which is identified through the leading generalized eigenspace of measurable matrices. Based on this result, addressing the best subset selection problem is suggested from a new perspective by solving an empirical generalized eigenvalue problem to estimate the direction of the high-dimensional vector of parameters. A new estimator is then studied based on the RIFLE algorithm and demonstrates a non-asymptotic bound of the L^2 risk, the minimax convergence of the estimator and a central limit theorem. Simulations show the superiority of the proposed inference over some known l_0 constrained estimators.

HO022 Room Amphitheater 2 MACHINE LEARNING AND TEXT-PROCESSING FOR HIGH-DIMENSIONAL DATA Chair: Miroslav Stefanik

H0206: **Estimating the number of entities with vacancies using administrative and online data**
Presenter: **Robert Pater**, University of Information Technology and Management, Poland
Co-authors: Maciej Beresewicz, Herman Cherniaiev

The number of entities is estimated to have at least one job vacancy. To achieve this goal, an alternative approach is proposed to the methodology exploiting survey data, which is based solely on data from administrative registers and online sources and relies on dual system estimation (DSE). To achieve this, job offers collected from online job boards in Poland and administrative data from public employment services are used. As these sources do not cover the whole reference population and the number of units appearing in all datasets is small, a DSE approach is developed for negatively dependent sources. To achieve the main goal, a thorough data cleaning procedure is conducted in order to remove out-of-scope units, identify entities from the target population, and link them by identifiers to minimize linkage errors. The effectiveness and sensitivity of the proposed estimator are verified in simulation studies. From a practical point of view, the results show that the current vacancy survey in Poland underestimates the number of entities with at least one vacancy by about 10-15%. The main reasons for this discrepancy are non-sampling errors due to non-response and under-reporting, which is identified by comparing survey data with administrative data.

H0164: **Forecasting online job vacancy attractiveness**
Presenter: **Zuzana Kostalova**, Slovak Academy of Sciences, Slovakia
Co-authors: Miroslav Stefanik, Stefan Lyocsa

The purpose is to explore whether predictions of online job vacancies (OJVs) attractiveness by job seekers, measured by i) number of job ad views, ii) response, and iii) conversion rate (the ratio of the two), could be improved. Apart from standard machine learning models, network-based feature extraction methods are used. Forecasting models utilize above 175 explanatory variables related to job characteristics, prerequisites and benefits, including simple textual features and even calendar effects. The approach could suggest what kind of data leads to the highest marginal contribution in forecasting OJV attractiveness. The findings could help employers better target prospective applicants and could be implemented in the search interface by the job portals to improve job matching, i.e., lead to improved recommender systems.

H0181: **Graph convolutional networks for bankruptcy prediction in P2P Bondora market**
Presenter: **Tomas Plihal**, Masaryk University, Czech Republic
Co-authors: Oleg Deev

The accurate prediction of bankruptcy in peer-to-peer (P2P) lending markets is a critical endeavour, yet current machine learning models often consider only individual attributes without adequately accounting for relational information among borrowers. The aim is to utilize a novel approach that leverages graph convolutional network (GCN) models for predicting bankruptcies in the P2P Bondora market. The proposed model incorporates borrower-specific features, such as credit history and loan information, and enriches this data by exploiting the high-order relational structures among borrowers through graph networks. The choice of GCN is motivated by its efficacy in capturing localized node features and edge connections, thus providing a comprehensive understanding of both node attributes and the graph topology. To validate the approach, its performance is compared against traditional classification models. The model offers insights into the relative importance of borrower attributes and network features, thereby contributing to the understanding of risk factors in P2P lending markets. By synthesizing node-specific attributes with graph-structured data, the model provides a more nuanced and effective tool for risk assessment. The findings have broad implications for enhancing decision-making processes in P2P lending platforms and offer avenues for future research in integrating graph theories and financial risk modelling.

H0182: **Forecasting financial cycle: Machine learning approach**

Presenter: **Stefan Lyocsa**, Slovak Academy of Sciences, Slovakia

Financial cycles are assumed to reflect the dynamics and interconnectedness between the credit, housing and stock markets, which are all important components of the overall financial stability. Estimates and accurate financial cycle forecasts could be useful for sound macro-prudential policy making and investment planning. We estimate the financial cycle for Slovakia and use machine learning techniques to predict 3- and 6-month-ahead levels of the financial cycle. The prediction accuracy is compared across multiple models and driven by a set of 170 potential predictors, including indicators related to banks, financial market, monetary policy, labor market, economic activity, business and consumer confidence and calendar effects.

Saturday 23.03.2024 16:10 - 18:15 Parallel Session D – HiTECCoDES2024

HI004 **Room Amphitheater 1** STATISTICAL MODELLING AND CAUSAL INFERENCE FOR TEXT DATA **Chair: Andrej Srakar**

H0160: **Optimal experimental designs for the industry: Case studies**
Presenter: **Kalliopi Mylona**, King's College London, United Kingdom

Applications of the optimal design methodology are demonstrated in complex pharmaceutical experiments. The aim was to explore the response surface with respect to various experimental factors, as well as to provide good-quality predictions in the experimental region. Optimality criteria corresponding to the fitted model inference and prediction quality are incorporated. The choice of the design that was used to run each experiment is discussed, and some interesting results have been obtained.

H0183: **Clustered Mallows model**
Presenter: **Luiza Piancastelli**, University College Dublin, Ireland
Co-authors: Nial Friel

Rankings are a type of preference elicitation that arises from experiments where judges are asked to arrange objects in decreasing order of utility. Orderings of an item set $\{1,\ldots,n\}$ yield permutations that reflect strict preferences amongst the objects. For various reasons, strict relations can be unrealistic assumptions in practical situations. One example is (I): the case that alternatives share common traits and henceforth could easily be indistinguishable. With moderate or large n, it becomes likely that evaluators are indifferent to some of their choices. Another possibility (II) is that, depending on the experiment, there can be a different importance attribution to the choices that form the rank. For example, judges could be mostly concerned with demonstrating their top and disfavored alternatives. In this top/bottom elicitation, middle-rank items could be close to the uniform placement of those remaining. This extends the famous Mallows model to accommodate indistinguishability, such as those in (I) and (II). The underlying groupings of items/choices within these scenarios motivate the naming clustered Mallows model (CMM). In addition to providing the flexibility to mix strict and indifference preferences, the CMM can also serve as a simplified representation of ranking data under large item sets.

H0211: **Spectral CLTs with long memory for causal inference in augmented large language models**
Presenter: **Andrej Srakar**, Institute for Economic Research Ljubljana, Slovenia

Since the pioneering works from the 1980s, central and noncentral limit theorems have been constantly refined, extended and applied to an increasing number of diverse situations. A recent study extended this to spectral central limit theorems valid for additive functionals of isotropic and stationary Gaussian fields. Their work uses the Malliavin-Stein method and Fourier analysis techniques for situations where Y_t admits Gaussian fluctuations in a long memory context. Another recent article augmented existing language models with long-term memory. They proposed a framework of language models augmented with long-term memory, which enables LLMs to memorize long histories. The two perspectives are combined with a CausalNLP, a toolkit for inferring causality with observational data that includes text in addition to traditional numerical and categorical variables, to develop spectral central limit theorems in a context of causality for text data from long-term memory augmented large language models. The main stochastic calculus tools are derived from the Malliavin-Stein method, Fourier analysis, and free probability. Applications on datasets are presented from finance and medical imaging. In conclusion, possible Bayesian extensions are discussed.

H0216: **Matching with text data: An experimental evaluation of methods for matching documents and of measuring match quality**
Presenter: **Reagan Mozer**, Bentley University, United States
Co-authors: Luke Miratrix

Matching for causal inference is a well-studied problem, but standard methods fail when the units to match are text documents: the high-dimensional and rich nature of the data renders exact matching infeasible, causes propensity scores to produce incomparable matches, and makes assessing match quality difficult. A framework for matching text documents is characterized that decomposes existing methods into (1) the choice of text representation and (2) the choice of distance metric. It investigates how different choices within this framework affect both the quantity and quality of matches identified through a systematic multifactor evaluation experiment using human subjects. Altogether, over 100 unique text-matching methods are evaluated, along with five comparison methods taken from the literature. The experimental results identify methods that generate matches with higher subjective match quality than current state-of-the-art techniques. The precision of these results is enhanced by developing a predictive model to estimate the match quality of pairs of text documents as a function of the various distance scores. The model was found to successfully mimic human judgment and also allows for approximate and unsupervised evaluation of new procedures in the context. The identified best methods are then employed to illustrate the utility of text matching in two applications.

HO020 **Room Amphitheater 2** RECENT DEVELOPMENTS IN MODELS FOR DATA PROCESSING **Chair: Marija Cuparic**

H0167: **Extrapolation before imputation reduces bias when imputing heavily censored covariates**
Presenter: **Sarah Lotspeich**, Wake Forest University, United States

Modelling symptom progression to identify informative subjects for a new Huntington disease clinical trial is problematic since time to diagnosis, a key covariate, can be heavily censored. Imputation is an appealing strategy where censored covariates are replaced with their conditional means, but existing methods saw over 200% bias under heavy censoring. Calculating these conditional means well requires estimating and then integrating the survival function of the censored covariate from the censored value to infinity. Existing methods use the semiparametric Cox model with Breslow's estimator to estimate the survival function flexibly. Then, for integration, the trapezoidal rule is used, but the trapezoidal rule is not designed for improper integrals and leads to bias. Calculating the conditional mean is proposed with adaptive quadrature instead, which can handle the improper integral. Yet, even with adaptive quadrature, the integrand (the survival function) is undefined beyond the observed data, so the Weibull extension is identified as the best method to extrapolate and then integrate. In simulation studies, it is shown that replacing the trapezoidal rule with adaptive quadrature and adopting the Weibull extension corrects the bias seen with existing methods. It further shows how imputing with corrected conditional means helps prioritize patients for future clinical trials.

H0208: **Testing independence for spherical and hyperspherical data: Kernel-based approach**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Marija Cuparic, Bruno Ebner

In diverse applied research areas, encountering spherical and hyperspherical data is common, highlighting the essential task of assessing independence within such data structures. In this context, some properties of test statistics that rely on distance correlation measures initially introduced for energy distance are presented, and their generalizations are based on strongly negative definite kernels. One significant advantage of this method is its versatility across different types of directional data, allowing for the examination of independence among vectors of varying characteristics. In addition, they are shown to be powerful compared to existing competitors.

H0218: **Estimation and goodness-of-fit testing of Levy processes: The variance gamma process**
Presenter: **Gerrit Grobler**, North-West University, South Africa
Co-authors: Simos Meintanis, Emanuele Taufer, Denis Belomestny

The price of financial securities, such as an exchange rate or a stock, is typically modelled as a continuous time stochastic process. Due to empirical evidence of infrequent large movements of security prices, processes that allow for jumps are increasingly popular models to use. A family of continuous time stochastic processes that allow for jumps are exponential Levy processes that generalize the classical geometric Brownian motion process. However, choosing a specific model from the large family of Levy processes requires model validation. A goodness-of-fit test of the variance gamma process that utilizes the analytical tractability of its characteristic function is presented. Since the estimation of this model is required to apply the test, estimation methods based on a likelihood approach as well as on the characteristic function will be discussed. In addition, the newly developed goodness-of-fit test will be applied to a variety of historically observed security prices.

H0212: **Testing independence in the presence of data missing completely at random**
Presenter: **Marija Cuparic**, University of Belgrade, Serbia
Co-authors: Danijel Aleksic, Bojana Milosevic

The initial focus is on the general results related to the asymptotic properties of non-degenerate U-statistics when the data are missing completely at random. Then, the focus is on the problem of testing independence using the estimator of Kendall's Tau. Specifically, limiting results are provided when employing several commonly used imputation methods. In addition, the results of empirical power studies are summarized, and directions for further research are presented.

HO021 **Room Lecture room 1** **THEORETICAL AND COMPUTATIONAL ASPECTS OF COMPLEX DATA MODELS** **Chair: Matus Maciak**

H0220: **Financial time series modelling using artificial intelligence**
Presenter: **Michaela Matouskova**, Technical University of Liberec, Czech Republic
Co-authors: Jan Picek, Petr Prucha

The rapid advancements in artificial intelligence (AI) have showcased its powerful capabilities, leading to investigations of the efficacy of AI-powered forecasting algorithms compared to traditional methods. The focus is on comparing the performance of Prophet and Merlion, two AI-powered algorithms, with time series modelling in EViews 13 software. Historical data on commercial real estate prices in the European market were used to evaluate the accuracy of each model's predictions by comparing them to actual price movements. This comparative analysis will contribute valuable knowledge to the domain of prediction algorithms, highlighting the potential advantages and limitations of AI-powered methods in the context of financial time series modelling.

H0210: **Fast and optimal inference for change points in piecewise polynomials via differencing**
Presenter: **Shakeel Gavioli-Akilagun**, London School of Economics, United Kingdom
Co-authors: Piotr Fryzlewicz

The problem of uncertainty quantification in change point regressions is considered, where the signal can be piecewise polynomial of arbitrary but fixed degree. Disjoint intervals are sought, which, uniformly at a given confidence level, must each contain a change point location. A procedure is proposed based on performing local tests at a number of scales and locations on a sparse grid, which adapts to the choice of the grid in the sense that by choosing a sparser grid, one explicitly pays a lower price for multiple testing. The procedure is fast as its computational complexity is always of the order $O(n\log(n))$ where n is the length of the data, and optimal in the sense that under certain mild conditions, every change point is detected with high probability and the widths of the intervals returned match the mini-max localization rates for the associated change point problem up to log factors. A detailed simulation study shows that the procedure is competitive against state-of-the-art algorithms for similar problems. The procedure is implemented in the R package ChangePointInference, which is available via GitHub.

H0173: **Bootstrapping not independent and not identically distributed data**
Presenter: **Martin Hrba**, Charles University, Praha, Czech Republic
Co-authors: Matus Maciak, Barbora Pestova, Michal Pesta

Classical normal asymptotics could bring serious pitfalls in statistical inference because some parameters appearing in the limit distributions are unknown and, moreover, complicated to estimate (from a theoretical as well as computational point of view). Due to this, plenty of stochastic approaches for constructing confidence intervals and testing hypotheses cannot be directly applied. Bootstrap seems to be a plausible alternative. A methodological framework for bootstrapping non-independent and not identically distributed data is presented together with a theoretical justification of the proposed procedures. Among others, bootstrap laws of large numbers and central limit theorems are provided.

H0176: **Selective pivot log-ratio coordinates for classification in high-dimensional compositional data**
Presenter: **Karel Hron**, Palacky University, Czech Republic
Co-authors: Nikola Stefelova, Julie de Sousa, Javier Palarea-Albaladejo, Dana Dobesova, Ales Kvasnicka, David Friedecky

Data from high-throughput biological experiments are often of a relative nature. This means that the most relevant information lies in the shape of the data distribution over the biological features rather than in the size of the measurements themselves. A well-established way to account for this in statistical processing is the log-ratio methodology of compositional data. Selective pivot log-ratio coordinates are introduced as a new type of orthonormal log-ratio coordinate representation for high-dimensional compositional data. This proposal aims to enhance the identification of biomarkers in the context of binary classification problems, which is a common setting of scientific studies in this field. These log-ratio coordinates are constructed such that the pivot coordinate representing a given compositional part aggregates all pairwise log-ratios of that part with the rest but, unlike in the usual formulation, excludes those that deviate from the main pattern. This novel coordinate system is embedded in a partial least squares discriminant analysis (PLS-DA) model for practical application. Using both synthetic and real-world metabolomic datasets, we demonstrate the enhanced performance of the novel approach compared to other methods used in the field.

H0179: **Online instability detection in a nonlinear expectile model: Theoretical and computational aspects**
Presenter: **Matus Maciak**, Charles University, Czech Republic

An automatic data-driven changepoint detection test is proposed to detect specific instabilities within a nonlinear regression framework. Conditional expectiles, well-known in econometrics for being the only coherent and elicitable risk measure, introduce additional 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, finite sample empirical properties and real data illustrations are presented.

Sunday 24.03.2024 09:00 - 10:15 Parallel Session F – HiTECCoDES2024

HO019 Room Amphitheater 1 STATISTICAL INFERENCE Chair: Andreas Artemiou

H0156: **Statistical inference for extreme value analysis**
Presenter: **Alexandros Karagrigoriou**, University of The Aegean, Greece
Co-authors: Ioannis Mavrogiannis, Ilia Vonta, Georgia Papasotiriou

Log-concavity and log-convexity play a key role in various scientific fields, especially in those where the distinction between exponential and non-exponential distributions is necessary for inferential purposes. A testing procedure for the tail part of a distribution, which can be used for the distinction between exponential and non-exponential distributions, is introduced. The conspiracy and catastrophe principles are initially used to establish a characterization of (the tail part of) the exponential distribution, which is one of the main contributions of the present work, leading the way for the construction of the new test of fit. The proposed test and its implementation are thoroughly discussed, and an extended simulation study has been undertaken to clarify issues related to its implementation and explore the extent of its capabilities. A real data case is also investigated.

H0158: **Statistical analysis in the presence of several competing risks**
Presenter: **Andreas Makrides**, University of the Aegean, Greece
Co-authors: Alexandros Karagrigoriou, Ilia Vonta, Theodora Dimitrakopoulou

The focus is on the description of the competing risks model in terms of the multi-state systems (MSS) methodology and the associated statistical inference when the sojourn times, i.e. the waiting times on each state, follow distributions belonging to a general class of distributions which is closed under minima.

H0169: **Addressing complex feature relationships: Harnessing the kernel association coefficient for nonlinear associations**
Presenter: **Kimon Ntotsis**, University of Leicester, United Kingdom
Co-authors: Andreas Artemiou, Alexandros Karagrigoriou

In the exploration of feature associations, conventional methods often assume linearity. However, real-world data frequently reveals non-linear relationships, necessitating innovative approaches for accurate assessment. The purpose is to introduce a novel kernel association coefficient designed to identify complex associations between features. Comparative analyses against existing coefficients consistently demonstrate higher accuracy. The proposed methodology proves effective across various data distributions and sample sizes, avoiding bias towards linear or non-linear associations. These findings contribute to the advancement of statistical modelling, emphasizing the importance of capturing intricate relationships for inferential, descriptive, or predictive purposes.

HO013 Room Amphitheater 2 ADVANCES IN STATISTICAL TOOLS FOR ECONOMICS AND FINANCE Chair: Alessandra Amendola

H0184: **Initial coin offerings: Can ESG mitigate underpricing?**
Presenter: **Alessandro Bitetto**, University of Pavia, Italy
Co-authors: Paola Cerchiello

Initial coin offerings (ICOs) have emerged as a novel way of start-up funding based on blockchain technology, and the aim is to explain the nexus between ICOs' success and underpricing, i.e. when the price of the offered token is lower than the one traded in the market. In particular, the focus is on the impact of the environmental, social and governance (ESG) pillars on the ICOs' underpricing. Therefore, a big and comprehensive dataset is built up, comprising 8000 ICOs spanning from 2015 through 2023, containing both technical and financial information. Moreover, an ESG score is assessed by means of advanced textual analysis performed over the whitepapers. The main results show that a higher ESG orientation leads to less underpricing, especially in the early trading days.

H0187: **Political cohesion and economic growth: The case of GDP for Italy**
Presenter: **Alessandro Grimaldi**, University of Salerno, Italy
Co-authors: Alessandra Amendola, Walter Distaso

The purpose is to derive new textual political polarity indices based on the text analysis of the entire collection of the Italian Senate of the Republic verbatim reports. The procedure allows for building a set of polarity indices reflecting the impact of the tone of political debate - as well as agreement/disagreement within political groups - on a specific economic variable over time. Time series regressions on the yearly Italian GDP growth rate point to a nontrivial predictive power of the proposed polarity indices, which, importantly - differently from common practice in related textual analysis literature - rely on a machine learning approach rather than the subjective choice of an affective lexicon.

H0191: **Examining the influence of ESG rates on the composition of the global minimum variance portfolio**
Presenter: **Vincenzo Candila**, University of Salerno, Italy
Co-authors: Luigi Aldieri, Alessandra Amendola

Recently, there has been a significant increase in attention to environmental, social, and governance (ESG) responsibilities. Corporations are now highly attuned to these issues, publicly disclosing not only traditional balance sheets but also environmental and sustainability reports. However, the question of how financial markets respond to strong ESG practices remains unanswered. Specifically, it is unclear whether investors show greater interest in portfolios of companies with high ESG ratings. The aim is to address this question, employing a comprehensive analysis based on a substantial panel of constituents (by weight) from the S&P500 index. Utilizing established multivariate techniques to calculate the conditional covariance matrix of the stocks under examination, the time-varying weights of the global minimum variance portfolio are determined. The uniformity of these weights across low, medium, and high ESG ratings provides insights into the relatively limited emphasis placed by financial markets on the ESG pillars, at least up to the present moment.

HP001 Room Lecture room 1 POSTER SESSION (VIRTUAL) Chair: Marios Demosthenous

H0196: **Causal models and phylo-spatio-temporal, multidimensional epidemiology: Dark figure estimation**
Presenter: **Andrzej Jarynowski**, FU Berlin, Germany
Co-authors: Vitaly Belik

Triangulation of data and methods is the way to get into findings (tools previously developed by statisticians and econometrists) to offer valuable insights for ONE public health. A multidimensional analysis was conducted to understand disease transmission in Poland, integrating key branches. A) A/H5N1 Epizootic in Cats: i) In-depth analysis of the A/H5N1 Epizootic in cats, involving positive and negative cases, RNA sequences, and participatory epidemiology data. ii) Utilization of daily time series data, revealing phylodynamic spatiotemporal clusters, patterns of disease spread, and hotspots of transmission. iii) Identification of separate virus introductions in eastern and western Poland, with A/H5N1 likely circulating in cats a month before the first confirmed case. B) Impact of healthcare access on COVID-19 Burden: i) Causal modelling to analyze the relationship between healthcare access and COVID-19 incidence. ii) BIC to select the optimal structure of the model (paths), nonparametric bootstrap to assess the strength of links in the model. iii) Highlighting the significant role of healthcare access in shaping geographical variations in COVID-19 burden. iv) Providing a nuanced understanding of healthcare access' contribution to unreported or undiagnosed COVID-19 infections and the effect of vaccines on preventing deaths.

H0199: **Analyzing the volatility of daily time series FOREX rates by using GARCH models: A case study from Albania**

Presenter: **Altin Kulli**, Qiriazi University College, Albania

Time series properties and predictability of exchange rates in transition economies considered. A period 2009-2011, prior euro daily Albanian exchange rates, with respect to four major currencies (USD, DEM, drachma and the Italian lira), were considered. The forex movements, as the most important factors affecting sales, profit forecasts, capital budgeting plans and international investment value, are of particular significance for the Albanian economy as an emerging country. The time-varying characteristics of forex volatility were explored using the GARCH methodology, discovering that the GARCH models might adequately describe the conditional second moments of these exchange rates of Albanian LEK; furthermore, upon consideration of past information, the percentage changes of such FOREX rates be predictable. The first lagged percentage changes for dollar and drachma were both significant and approximately the same, which is to be expected for an emerging economy, indicating market inefficiency. The volatility is well represented by a GARCH (1, 1) (it might be predictable on past innovations basis, volatility measures). The conditional variance shock did not persist and quickly died out. The weekend effect on trade opening was found to be positive for all except the USD series, which had its value equal to zero. The volatility has increased from Friday to Monday for all, mainly for Mark and Drachma.

Sunday 24.03.2024 10:45 - 12:25 Parallel Session G – HiTECCoDES2024

HI025 Room Amphitheater 2 TEXT MINING Chair: Cristian Gatu

H0165: **Analyzing the impact of removing infrequent words on topic quality in LDA models**
Presenter: **Viktoriia Naboka-Krell**, Justus Liebig Unversity of Giessen, Germany
Co-authors: Victor Bystrov, Anna Staszewska-Bystrova, Peter Winker

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 the removal of infrequent terms 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.

H0193: **Automatic detection of industry sectors in legal articles using machine learning approaches**
Presenter: **Stella Hadjiantoni**, University of Essex, United Kingdom
Co-authors: Berthold Lausen, Hui Yang, Ruta Petraityte, Yunfei Long

The ability to automatically identify industry sector coverage in articles on legal developments, or any kind of news articles for that matter, can bring plentiful benefits both to the readers and the content creators themselves. By having articles tagged based on industry coverage, readers would be able to get to legal news that is specific to their region and professional industry. A machine learning-powered industry analysis approach which combined natural language processing (NLP) with machine learning (ML) techniques was investigated. A dataset consisting of over 1,700 annotated legal articles was created for the identification of six industry sectors. Text and legal-based features were extracted from the text. Both traditional ML methods (e.g. gradient boosting machine algorithms and decision-tree based algorithms) and deep neural networks (e.g. transformer models) were applied for performance comparison of predictive models. The system achieved promising results with area under the receiver operating characteristic curve scores above 0.90 and F-scores above 0.81 with respect to the six industry sectors. The experimental results show that the suggested automated industry analysis, which employs ML techniques, allows the processing of large collections of text data in an easy, efficient, and scalable way. ML methods perform better than deep neural networks when only a small and domain-specific training data is available for the study.

H0195: **Data augmentation for testing subject alignment with a COST Action**
Presenter: **Louisa Kontoghiorghes**, Kings College London, United Kingdom
Co-authors: Ana Colubi

The objective is to use a generative pre-trained transformer (GPT) model for data augmentation when limited text data is observed. The focus is to assess how well a research abstract aligns with the scientific objectives of a specific COST action. To achieve this, a GPT model is employed to create abstracts based on the proposal of the COST action, ensuring contextual relevance. The model is also used to generate a variation set of abstracts from the one of interest. To quantify differences between the two sets, Latent Dirichlet Allocation (LDA), a topic model method, is implemented, the prevalence is estimated, and a two-sample bootstrap test is performed, providing a statistical comparison of subject alignment.

H0150: **Weighted degrees and truncated derived networks**
Presenter: **Vladimir Batagelj**, IMFM, Slovenia

Large bibliographic networks are sparse - the average node degree is small. This is not necessarily true for their product - in some cases, it can "explode" (it is not sparse, increases in time and space complexity). An approach in such cases is to reduce the complexity of the problem by limiting the attention to a selected subset of important nodes and computing with corresponding truncated networks. The nodes can be selected by different criteria. An option is to consider the most important nodes in the derived network - nodes with the largest weighted degree. It turns out that the weighted degrees in the derived network can be computed efficiently without computing the derived network itself.

HO023 Room Amphitheater 1 HIGH DIMENSIONAL STATISTICS Chair: Andreas Artemiou

H0168: **Generalized sufficient dimension reduction in the presence of categorical predictors**
Presenter: **Ben Jones**, Aerospace Sector, United Kingdom

Measure-theoretic developments in sufficient dimension reduction have enabled its application with predictors and/or responses lying in separable metric spaces, while allowing nonlinear reductions. A significant limitation of these developments is that they do not allow for the presence of additional categorical predictors, which we want to use to constrain the dimension reduction. An extension which overcomes this limitation is presented. Conceptual definitions are first given to set up the problem technically, then the novel estimator "partial generalized sliced inverse regression" of the target of estimation is described. The results of this method are further illustrated by real-world data.

H0186: **Distributed estimation and inference in sparse conditional Gaussian graphical models under an unbalanced setting**
Presenter: **Eugen Pircalabelu**, Universita catholique de Louvain, Belgium

The focus is on a distributed estimation and inferential framework for sparse multivariate regression and conditional Gaussian graphical models under the unbalanced splitting setting. This type of data splitting arises when the datasets from different sources cannot be aggregated on one machine or when the available machines are of different powers. The number of covariates, responses and machines grows with the sample size while sparsity is imposed. Debiased estimators of the coefficient matrix and of the precision matrix are proposed on every single machine, and theoretical guarantees are provided. Moreover, new aggregated estimators that pool information across the machines using a pseudo-log-likelihood function are proposed, and it is shown that they enjoy consistency and asymptotic normality as the number of machines grows with the sample size. The performance of these two estimators is investigated via a simulation study and a real data example. Empirically, it is shown that the performances of these estimators are close to those of the non-distributed estimators that use the entire dataset.

H0189: **A method for sparse and robust independent component analysis**
Presenter: **Lauri Heinonen**, University of Turku, Finland
Co-authors: Joni Virta

Independent component analysis (ICA) is a popular family of methods for decomposing signals into independent sources. One group of ICA methods are those achieved by using symmetrized scatter matrices or other scatter matrices with independent properties in invariant coordinate selection (ICS). A sparse version of this method is presented to achieve sparse independent component analysis (SICA). When using scatter matrices that are also robust, the SICA method is also robust. Compared to regular ICA, sparse ICA gives sparse loadings where some of the loadings are estimated to be exactly zero. The SICA method is presented as a sequence of regression problems, and the LASSO penalty is used to achieve sparsity. The method is illustrated with examples and compared to different ICA methods and other relevant competitors.

H0223: **Envelope-based support vector machine classifier**
Presenter: **Alya Alzahrani**, Taif University, Saudi Arabia

The envelope method is a relatively new and efficient dimension reduction technique. We extended this method to classification and developed a new projection-based approach based on a Support Vector Machine (SVM) classifier. The proposed classifier is obtained by combining the envelope method and SVM to achieve a better and more efficient classification. Using the idea of the envelope to extract a lower-dimensional subspace projected the data on has advanced the classification performance.

HO024 Room Lecture room 1 STATISTICAL MODELS AND METHODS FOR EDUCATION EVALUATION Chair: Marialuisa Restaino

H0180: **The role of schools in shaping university careers: Evidence from Italy**
Presenter: **Cristian Usala**, University of Cagliari, Italy
Co-authors: Mariano Porcu, Isabella Sulis

The aim is to investigate how schools influence students' careers at the university by focusing on their academic status at the beginning of their second year of careers. High school's impact is modelled by using the MOBYSU.IT database on the population of students enrolled in an Italian university between 2015 and 2018. In particular, a two-step approach is applied to account for the role played by other confounding factors related to students' characteristics, disciplinary fields, and socio-economic conditions of the areas. The first step entails two fixed-effects regressions to estimate the average effect of schools and disciplinary fields on the number of credits earned by students during their first year. This step derives two indirect indicators: one measuring schools' effectiveness in providing students with the necessary competencies and one providing insights into the difficulty level of the chosen field and program. In the second step, a multinomial logit model is estimated to evaluate the effect of schools on students' probability of being regulars, at risk of dropout, dropouts, and changing degree programs and/or universities while controlling for a wide set of covariates. The results show a significant and positive school effect on students' university careers.

H0192: **What we learn from PISA 2022?**
Presenter: **Mariangela Zenga**, Universita degli Studi di Milano-Bicocca, Italy
Co-authors: Adele Marshall

A statistical analysis is proposed for the assessment on education systems, with a specific focus on analyzing PISA 2022 data. PISA, known as the Programme for International Student Assessment, is a global benchmark for evaluating the performance of 15-year-old students across several countries. The focus will be on the examination of student proficiency in mathematics, reading, and science, as reported in the most recent PISA dataset. By utilizing multilevel analysis, the objective is to uncover insights into the educational landscape across several countries. Through this comprehensive exploration, the aim is to identify underlying patterns, trends, and challenges that influence students' learning outcomes.

H0194: **Clustering of Italian higher education institutions based on a destination: Specific approach**
Presenter: **Luca Scaffidi Domianello**, University of Catania, Italy
Co-authors: Silvia Bacci, Bruno Bertaccini

Student mobility flows are typically examined using gravity models. These models generally assume a uniform relationship for each origin-destination pair. However, in cases where spatial interaction behaviour varies across the space, the estimated parameters reflect an average of these different relationships. This assumption may not hold for the Italian higher education system, where, as a consequence of the decentralization process, some universities have a national vocation while others primarily target local populations. For this reason, a destination-specific approach is adopted to gather detailed insights for each university. The empirical analysis based on destination-specific models reveals distinct interaction dynamics among Italian universities, corroborating the hypothesis of spatial heterogeneity. Additionally, a fuzzy clustering technique is proposed based on estimated distance parameters that measure the deterrence effect. The underlying idea is that universities with lower distance parameters are associated with wider catchment areas. Through this clustering procedure, two distinct groups are identified: one comprising universities with a national vocation and the other comprising those with a local focus.

H0201: **Educational data mining for predicting students' success**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Marcella Niglio, Michele La Rocca, Maria Prosperina Vitale

Educational Data Mining (EDM) is an emerging research field that focuses on the application of techniques and methods of data mining in educational environments. The focus is on "student success", intended as the ability of students to close a given educational level successfully. It is a crucial element of evaluation, and it is often used as a criterion to assess the quality and performance of educational institutions. Early detection of the "students at risk" (with a high probability of dropping out of the educational institution) and the adoption of preventive measures can help decision-makers to provide and plan proper actions for improving students' performances (and consequently their success), and eventually revise the educational project. The aim is to explore the main differences in students' performance among bachelor's degrees by using regression models. The analysis concerns students enrolled at 3-year degrees in an Italian university (located in the South of Italy) during ten academic years. Student success is measured in terms of the number of ECTS credits earned during the first year. Hence, the main purposes are to i) estimate the probability of getting at least a certain number of credits at the end of the first year, ii) identify which students' features might affect it, and iii) classify students according to their churn risk.

Sunday 24.03.2024 14:00 - 15:40 Parallel Session H – HiTECCoDES2024

HO012 Room Amphitheater 1 LOW-DIMENSIONAL STRUCTURES IN HIGH-DIMENSIONAL DATA Chair: Dan Vilenchik

H0207: **The DNA of sarcasm and its implication in cross-domain tasks**
Presenter: **Havana Rika**, The Academic College of Tel Aviv - Yaffo, Israel
Co-authors: Dan Vilenchik, David Ben-Michael

Sarcasm is a form of figurative language where the speaker intends to convey a message implicitly. The explicit meaning of a sarcastic statement is often contradictory to the implicit meaning and heavily reliant on the context. In some instances, the sarcastic intent may be accentuated by the speaker's tone of voice, which is absent in the written text. As a result, detecting sarcasm is a non-trivial task for humans, let alone for automatic methods. The problem of sarcasm detection has traditionally been approached as a binary classification task, aiming to predict whether a given text contains sarcasm or not. In recent years, there has been a growing trend to address sarcasm detection using deep neural network (DNN) models. These models have been solely evaluated through the in-domain method and, unfortunately, demonstrate poor performance in the cross-domain evaluation method (training on one and testing on another). The purpose is to explain the low cross-domain performance through the many shades of sarcasm. For example, some remarks are more humorous than others, while others may be more toxic; in short, not all sarcastic comments were born equal. The differences are identified and presented among a variety of well-known sarcasm datasets. Using these insights, a data enrichment procedure is guided that significantly improves cross-domain performance up to an additive 13% in F1 score without requiring labelled data.

H0203: **Towards reverse algorithmic engineering of neural networks**
Presenter: **Dan Vilenchik**, Ben-Gurion University, Israel

As machine learning models get more complex, they can outperform traditional algorithms and tackle a wider range of problems, including challenging combinatorial optimization tasks. However, this increased complexity can make understanding how the model makes its decisions difficult. Explainable models can increase trust in the models' decisions and may even lead to improvements in the algorithm itself. Algorithms like GradCAM or SHAP provide good explanations in terms of feature importance, typically for classification tasks, but they provide little insight when the ML pipeline is designed to work, for example, as an algorithm for solving optimization problems. A framework for explaining a model's decision-making process is presented from an algorithmic point of view while taking into account domain knowledge of the problem at hand. Using the NeuroSAT algorithm for SAT solving as a case study, it is demonstrated how the framework explains the underlying algorithmic concepts that drive the operation of an NN-based model. For example, it is discovered that for sparse random SAT instances, NeuroSAT mimics the pure literal heuristic, while for denser formulas, it relies on the concept of support to decide which variables to flip, similar to the WalkSAT algorithm.

H0213: **Drug repurposing using link prediction on knowledge graphs**
Presenter: **Sarel Cohen**, The Academic College of Tel Aviv-Yaffo, Israel

The active global SARS-CoV-2 pandemic caused more than 426 million cases and 5.8 million deaths worldwide. The development of completely new drugs for such a novel disease is a challenging, time-intensive process. Despite researchers around the world working on this task, no effective treatments have been developed yet. This emphasizes the importance of drug repurposing, where treatments are found among existing drugs that are meant for different diseases. A common approach to this is based on *knowledge graphs*, which condense relationships between entities like drugs, diseases and genes. Graph neural networks (GNNs) can then be used for the task at hand by predicting links in such knowledge graphs. Expanding on state-of-the-art GNN research, a recent study developed the Dr-COVID model. Their work using additional output interpretation strategies is further extended. The best aggregation strategy derives a top-100 ranking of 8,070 candidate drugs, 32 of which are currently being tested in COVID-19-related clinical trials. In addition, the implementation of the Dr-COVID model is improved by significantly shortening the inference and pre-processing time by exploiting data parallelism.

H0175: **Addressing high-dimensionality for dynamic principal components analysis in the frequency domain**
Presenter: **David Paul Suda**, University of Malta, Malta
Co-authors: Matthew Attard, Fiona Sammut, Dan Vilenchik

Dynamic principal components analysis refers to a class of dimension reduction methods of multivariate data in a time series setting. These methods are important as they address the handling of time-dependence and/or short-term correlation, which classical principal components analysis does not cater for. Brillinger's frequency domain approach is the earliest of such approaches and is aimed at a single realisation setting. In the last decade, time-domain approaches have also evolved, addressing both the single realisation and multiple realisation settings. Authors of the latter two approaches have also introduced a sparsity extension, which allows one to generalise these approaches to the high-dimensional data setting. Peer-reviewed literature addressing high-dimensionality in the frequency domain setting remains missing. The frequency domain approach to principal components essentially replicates the classical approach but on cross-spectra instead of the covariance matrix, ultimately recuperating the loadings through the Fourier inverse and the principal components through the dynamic Karhunen-Loeve expansion. The aim is to address the void concerning high-dimensionality in academic literature when it comes to frequency-domain principal components. This can be done by applying techniques for addressing sparsity in the static case, which can be readily adapted to the frequency-domain approach. Some preliminary results on real or simulated data are expected to be presented.

H0224: **A general framework for learning-augmented online allocation**
Presenter: **Ilan Cohen**, Bar Ilan University, Israel, Israel

Online allocation is a broad class of problems where items arriving online have to be allocated to agents who have a fixed utility/cost for each assigned item so to maximize/minimize some objective. This framework captures a broad range of fundamental problems such as the Santa Claus problem, Nash welfare maximization (maximizing geometric mean of utilities), makespan minimization (minimizing maximum cost), minimization of ℓ_p-norms, and so on. Even for divisible items, these problems are characterized by strong super-constant lower bounds in the classical worst-case online model. We study online allocations in the learning-augmented setting, i.e., where the algorithm has access to some additional (machine-learned) information about the problem instance. We introduce a general algorithmic framework for learning-augmented online allocation that produces nearly optimal solutions for this broad range of maximization and minimization objectives using only a single learned parameter for every agent. As corollaries of our general framework, we improve prior resultsfor learning-augmented makespan minimization, and obtain the first learning-augmented nearly-optimal algorithms for the other objectives such as Santa Claus, Nash welfare, ℓ_p-minimization, etc. We also give tight bounds on the resilience of our algorithms to errors in the learned parameters, and study the learnability of these parameters.

HO014 Room Amphitheater 2 COMPLEX DATA ANALYSIS: MODEL SPECIFICATION Chair: Bojana Milosevic

H0157: **Optimal testing for symmetry on the torus**
Presenter: **Sophia Loizidou**, University of Luxembourg, Luxembourg
Co-authors: Andreas Anastasiou, Christophe Ley

Several complex real-world data can be viewed as points on the hyper-torus, which is the cartesian product of circles. Over the past few years, this has motivated new proposals of distributions on the torus, both (pointwise) symmetric and sine-skewed asymmetric. In practice, it is relevant to know whether one should use the simpler symmetric models or the more convoluted yet more general asymmetric ones. So far, only parametric likelihood ratio tests have been defined to distinguish between a symmetric density and its sine-skewed counterpart. A new semi-parametric test is presented, a test which is valid not only under a given parametric hypothesis but also under a very broad class of symmetric distributions. A description of its construction and asymptotic properties under the null and alternative hypotheses will be presented. Using Stein's method, bounds for the rate of convergence of the test statistic are derived, and finite sample behavior (through Monte Carlo simulations) will be given, as well as an application of the test on protein data.

H0174: **New estimators for directional data**

Presenter: **Adrian Fischer**, Universita libre de Bruxelles, Belgium

Co-authors: Robert Gaunt, Yvik Swan

In Stein's method, one can characterize probability distributions with differential operators. These characterizations are used to obtain new point estimators for the parameters of spherical distributions. As a consequence of the usually simple form of the operator, explicit estimators are obtained in cases where standard methods, such as maximum likelihood estimation, require a numerical procedure to calculate the estimate. Among others, competitive estimators are obtained for the concentration parameter of the Fisher-von Mises distribution and an explicit estimator for all parameters of the Fisher-Bingham distribution.

H0202: **Multivariate sign tests for sphericity: Dealing with skewness and dependent observations**

Presenter: **Gaspard Bernard**, University of Luxembourg, Luxembourg

The problem of testing for the sphericity of a shape parameter is considered when the sample is drawn from a distribution with elliptical directions. This setting encompasses both cases where some skewness is present and cases where the i.i.d. hypothesis does not hold anymore. In the elliptical directions setting, the existing sphericity test based on the multivariate signs of the observations is valid if the location parameter is specified when constructing the multivariate signs. In practice, the location parameter needs to be estimated, and the asymptotic validity of the test will depend on the asymptotic cost of this estimation. This asymptotic cost is studied and shown under what conditions the test based on the multivariate signs is asymptotically valid and optimal.

H0217: **On estimating the mode of an angular distribution**

Presenter: **Jaco Visagie**, North-West University, South Africa

Co-authors: Fred Lombard, Charl Pretorius

Classes of estimators are proposed for the mode of an angular distribution, each adapted from a corresponding class of estimators defined on the real line. In addition to point estimation, the construction of confidence intervals using the bootstrap is considered. The asymptotic properties of some of the proposed estimators are outlined, and a Monte Carlo study is included in order to compare the finite sample performance of the proposed estimators.

Sunday 24.03.2024 16:10 - 17:25 Parallel Session I – HiTECCoDES2024

HI026 Room Amphitheater 1 FINANCIAL ECONOMETRICS Chair: Leopold Soegner

H0159: **Open-end monitoring of structural breaks in the cointegration VAR**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner

An open-end consistent monitoring procedure is developed with the goal of detecting structural changes in a Johansen-type error correction model. An open-end monitoring approach developed for the stationary case is adapted to the non-stationary case. This allows the investigation of breaks where the cointegration rank changes as well as breaks where the cointegration rank remains constant but the model parameters change. Non-parametric and parametric monitoring procedures are developed, where the test statistic is either based on moments or the model parameters.

H0166: **Extracting insights from large and complex datasets: Examples of dimensionality reduction by applying economic theory**
Presenter: **Tsvetomira Tsenova**, Experian Bulgaria, Bulgaria

Currently, an increased number of datasets containing individual microdata from surveys and regulatory reports of banks become publicly available, which increases the information universe for academics, policymakers and the general public. However, insights generation is insufficient due to the dataset's volume, complex dimensionality and changing structure over time. The lack of an adequately long and consistent time series structure hinders purely empirical research explorations. The purpose is to provide several examples of how economic theory could be used to enrich micro-data sets with additional statistical data and focus on certain dimensions for answering specific policy and general public questions. The examples include the survey of professional forecasters for the Euro Area and the United States, as well as EU regulatory bank balance sheet data. The insights relate to monitoring the state of inflation expectations, economic growth prospects, structural uncertainty, lending decisions, credit risk and financial stability.

H0170: **Gamma-driven Markov processes with application to realized volatility**
Presenter: **Wagner Barreto-Souza**, University College Dublin, Ireland
Co-authors: Fernanda Mendes, Sokol Ndreca

A novel class of Markov processes is proposed for dealing with continuous positive time series data, which is constructed based on a latent gamma effect and named gamma-driven (GD) models. The GD processes possess desirable properties and features: (i) it can produce any desirable invariant distribution with support on R+, (ii) it is time-reversible, and (iii) it has the transition density function given in an explicit form. Estimation of parameters is performed through the maximum likelihood method combined with a Gauss Laguerre quadrature to approximate the likelihood function. The evaluation of the estimators and also confidence intervals of parameters are explored via Monte Carlo simulation studies. Two generalizations of the GD processes are also proposed to handle non-stationary and long-memory time series. The proposed methodologies are applied to analyze the daily realized volatility of the FTSE 100 equity index.

HO016 Room Amphitheater 2 RISK ANALYSIS AND MACHINE LEARNING APPLICATIONS Chair: Robertas Alzbutas

H0198: **Machine learning and uncertainty analysis for remaining value estimation**
Presenter: **Ieva Dunduliene**, Kaunas University of Technology, Lithuania
Co-authors: Robertas Alzbutas

The estimation of the remaining value is gaining more attention, especially in the context of sustainability, engineering, and industry. Furthermore, the remaining value estimations present additional challenges related to uncertainties, subjective information integration, and exogenous feature dependencies. This task is nontrivial and complicated by the absence of a gold standard for estimating and comparing the calculated remaining values. The challenges especially arise while trying to estimate or/and validate the remaining value of the refurbished products. To address these challenges, it is proposed to integrate machine learning (ML) methods and uncertainty analysis to ensure the accuracy and risk minimization of the remaining value estimation. Machine learning methods are widely utilized in the remaining value estimation process due to the methods' ability to detect and identify relationships and hidden patterns among variables. Through case studies, the effectiveness of the presented approach is proven by providing interval estimates of the remaining value calculations that incorporate uncertainty assessment. This allows decision-makers and consumers to make well-informed decisions when considering the purchase of refurbished products. In conclusion, a framework is presented that integrates ML and uncertainty analysis for remaining value estimation by combining multiple ML models into one and facilitating the quantification of uncertainty in the results.

H0197: **Data clustering methods and large language models applications**
Presenter: **Mantas Lukauskas**, Kaunas University of Technology, Lithuania

The escalating complexity and volume of data in various scientific domains necessitate advanced methodologies for efficient data analysis and interpretation. The purpose is to delve into the synergy between data clustering methods and large language models (LLMs) to foster innovative approaches in handling extensive datasets. Data clustering, a pivotal aspect of data mining, involves grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups. It serves as a foundational step in the analysis, enabling the identification of intrinsic patterns and structures within the data. Simultaneously, the advent of LLMs, characterized by their vast parameter spaces and deep learning capabilities, has revolutionized natural language processing and understanding. It explores how these models can be leveraged to enhance the interpretability and applicability of clustering results, particularly in handling unstructured data, such as text. Through a synthesis of theoretical discussion and practical case studies, the presentation aims to highlight the potential gains of integrating clustering techniques with LLMs, offering insights into their applicability across various fields, from bioinformatics to social media analytics. This investigation not only broadens the understanding of data analysis methodologies but also opens avenues for future research in optimizing data handling and knowledge extraction processes.

H0200: **Delayed payment modelling using machine learning methods**
Presenter: **Mindaugas Kavaliauskas**, Kaunas University of Technology, Lithuania

Credit risk modelling holds particular significance for trade companies because it's common for them to permit the purchase of goods with deferred payment terms. Risk assessment models typically estimate a company's risk by using various financial metrics such as net profit, total revenue, working capital, total assets, and their corresponding ratios. One example of such a model is the Altman Z-score, a classic bankruptcy prediction model first published in 1968. However, these models have a significant drawback. These models are not well-suited for real-time risk assessment. Financial reports are typically published months after the end of the financial year. The aim is to utilize an alternative data - delayed invoice payment times for credit risk assessment. The attempt is to forecast future payment delays using past payment delay records. While this may seem like a typical time series forecasting problem, the data's nature is quite distinct: the time intervals between invoices are irregular, and the number of invoices for a particular company can vary from just one invoice in its history to over a dozen invoices per month. Building models based on this kind of data requires additional data preprocessing procedures such as padding, trimming, etc. Several data preprocessing methods are explored, and a few statistical and machine learning models are applied. The accuracy of these models is provided and discussed.

Authors Index