PROGRAMME AND ABSTRACTS

HiTEc meeting & Workshop on

Complex data in Econometrics and Statistics (HiTEc & CoDES 2025)

https://www.cmstatistics.org/hiteccodes2025

Cyprus University of Technology, Limassol, Cyprus
8-9 July 2025

Contents

Tuesday 08.07.2025 09:30 - 10:20 Room: Amphitheater 1 Chair: Erricos Kontoghiorghes Keynote talk I

Consistent estimation of linear regression models from different data sources with many variables in common

Speaker: **Masayuki Hirukawa, Ryukoku University, Japan**

When conducting regression analysis, econometricians often face situations where some regressors are unavailable in the primary dataset (e.g., an ability measure in wage regression). Suppose that they can find an auxiliary dataset that contains missing regressors as well as other variables common across two datasets (overlapping variables). Under this environment, it is possible to estimate regression coefficients consistently by combining primary and auxiliary datasets. Examples of such estimation procedures are the matched-sample indirect inference (MSII) and the plug-in least squares (PILS). However, these estimators can attain the parametric convergence rate only if the number of overlapping variables is three or less. Then, the scope of MSII and PILS is extended so that both can restore the parametric convergence rate when primary and auxiliary datasets have many overlapping variables. The extension takes three steps, namely, (i) dimension reduction under some structural assumption on the conditional expectations of missing regressors given overlapping variables, (ii) imputation of proxies for missing regressors, and (iii) estimation of the regression model. Convergence properties of extended MSII and PILS are explored in conjunction with covariance estimation. Monte Carlo simulations confirm their nice finite-sample properties, and a real data example of intergenerational income mobility is also presented.

Tuesday 08.07.2025 17:35 - 18:25 Room: Amphitheater 1 Chair: Demetris Koursaros Keynote talk II

The environmental impact of green bonds: Data-driven insights through statistical learning

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

Green bonds have become a transformative financial instrument aimed at financing initiatives with positive environmental outcomes. We investigate the efficacy of corporate green bond issuances in improving environmental performance at an individual issuer level, providing insights into issuer characteristics and factors influencing environmental impact. Based on a controlled interrupted time-series (CITS) model, green bonds and conventional bonds are compared to estimate the individual effects of green bonds on issuers' environmental (E) scores. This is followed by a two-stage statistical analysis. In the first stage, a random forest model is built to classify the key factors influencing environmental performance improvements, revealing that bond attributes, notably Climate Bonds Initiative certification, had minimal predictive power. In the second stage, we employ a generalized additive model to capture non-linear relationships between green bond effects and additional explanatory variables. Our results show that green bonds yield variable impacts on E scores across issuers. Non-linear relationships are observed, especially for bond value and issuer characteristics, affecting E-score changes. Despite some green bonds not affecting performance significantly, the study underlines the positive environmental impact of green bonds overall, urging policymakers to support green financing, yet highlighting issuer characteristics as more influential than the bonds themselves.

Wednesday 09.07.2025 17:35 - 18:25 Room: Amphitheater 1 Chair: Ana Colubi Keynote talk III

Detection of chaotic behaviors in stochastic systems by dispersion ratios in local block Lyapunov exponents diagrams

Speaker: **Yan Liu, Waseda University, Japan** Rintaro Ichiya, Rinka Sagawa

Statistical methods are presented for quantifying chaotic behaviors in stochastic processes. We introduce the local block Lyapunov exponent and the diagonal Lyapunov dispersion ratio as fundamental statistical tools to distinguish chaotic behaviors in stochastic processes. We develop the asymptotic theories for these statistical tools under a general setting. As a macroscopic measure, we quantify the sensitive dependence induced by chaos in each block, and investigate the distributional distortion in stochastic processes. Numerical simulations under different parameter settings illustrate the satisfactory performance of our statistical approach. We also apply this method to the financial market data, providing evidence for the possible chaotic behaviors in the data.

Tuesday 08.07.2025 10:50 - 12:30 Parallel Session B – HiTECCoDES2025

HI002 Room Amphitheater 1 HiTEc session: High-dimensonal and complex data Chair: Enea Bongiorno

H0162: **Low-rank and sparse network regression**
Presenter: **Weining Wang**, University of Bristol, United Kingdom

Spillover effects in spatial network models are analyzed under settings where measurement noises might contaminate the neighborhood (i.e., adjacency) matrix. We propose to adopt the low-rank and sparse structure to capture the stylized network pattern in empirical datasets. We develop a robust estimation framework via regularization techniques: the Least Absolute Shrinkage and Selection Operator (LASSO) for the sparse component and a nuclear norm penalty for the low-rank component. We propose two estimators: (1) A two-stage procedure that first de-noises the adjacency matrix via regularization and subsequently integrates the purified network to regression analysis, and (2) A single-step supervised Generalized Method of Moments (GMM) estimator jointly estimates the regression parameters and refines the network structure. Simulation evidence underscores the superiority of our approach. In scenarios with noisy networks, our method reduces the root mean squared error (RMSE) of coefficient estimates by 5070% compared to conventional GMM. This advantage is more significant when network contamination is endogenous, a common challenge in empirical settings where measurement errors correlate with the observed outcomes.

H0198: **On data subsampling for Poisson regression**
Presenter: **Alexander Munteanu**, TU Dortmund, Germany
Co-authors: Han Cheng Lie

The aim is develop and analyze new data subsampling techniques for Poisson regression with count data $y \in \mathbb{N}$. In particular, we consider the Poisson generalized linear model with ID and square-root link functions. We consider the method of *coresets*, small weighted subsets that approximate the log-likelihood up to a factor of $1 \pm \varepsilon$. By introducing a novel complexity parameter ρ and a domain shifting approach, we show that sublinear coresets with $1 \pm \varepsilon$ approximation guarantee exist when ρ is small. In particular, the number of input points can be reduced to polylogarithmic. We show that the dependence on other input parameters can also be bounded, though not always logarithmically. In particular, we show that the square root-link admits $O(\log(y_{\max}))$ dependence on the largest count, while the ID-link requires $\Theta(\sqrt{y_{\max}/\log(y_{\max})})$.

H0174: **Drilling into Erasmus learning mobility flows between countries 2014-2024**
Presenter: **Vladimir Batagelj**, IMFM, Slovenia

Alternative exploratory views on the network Erasmus+ learning mobility flows since 2014 are proposed. It has 35 nodes (countries), is very dense, and the range of link weights (number of visits) is very large (from 1 to 217003). A monotonically increasing transformation is used to reduce the range. Traditional graph-based visualization is unreadable. To gain insight into the structure of a dense network, it can be reduced to a skeleton by removing less essential links and/or nodes. We have determined a 1-neighbors and 2-neighbors subnetworks. The 1-neighbors skeleton highlights Spain as the main attractor in the network. The 2-neighbors skeleton shows the dominant role of Spain, Germany, France, and Italy. The Ps cores approach confirms these observations. Using the “right” order of the nodes in a matrix representation can reveal the network structure as block patterns in the displayed matrix. The clustering of network nodes based on Salton dissimilarity again shows the dominant role of Spain, Germany, France, and Italy, but also two main clusters of the West-East (developed-less developed) country division. The Balassa normalization (log(measured/expected) visits) matrix shows that most visits within the two main clusters are above expected, while most visits between them are below expected. `https://github.com/bavla/wNets/tree/main/Data/Erasmus`

H0165: **Estimating differential entropy in high dimensional spaces**
Presenter: **David Weston**, Birkbeck University of London, United Kingdom

Entropy estimation of a continuous distribution can be achieved by first estimating the density. However, such plug-in estimators often exhibit high variance, especially in high-dimensional spaces. A well-known graph-theoretic approach to reduce the variance and increase the bias in differential entropy estimation involves constructing a minimal spanning tree from the data, where the estimator is a function of the lengths of its edges. An alternative approach is demonstrated using a measure-preserving space-filling curve to induce a Hamiltonian. The reason why this approach results in a poor estimator is explained, along with a method to enhance its performance significantly. To demonstrate the utility of the proposed method, a simple classification approach is applied to image data, which serves as an example of high-dimensional data.

HO005 Room Amphitheater 2 Applied econometrics and ML in economic institutions Chair: Daria Scacciatelli

H0207: **Measuring economic sentiment in italian news**
Presenter: **Andrea Rollin**, Ministry of Economy and Finance Italy, Italy
Co-authors: Matteo Berta, Loredana Rinaldi, Chiara Eleonora De Marco, Maksim Bondarenko, Sara DAndrea, Ferroni Valeria, Valeria Macauda, Daria Scacciatelli, Salvatore Greco, Simone Monaco, Salvatore Lo Sardo, Daniele Apiletti, Tania Cerquitelli

Measuring economic sentiment in news articles offers a rapid window into public perception regarding the economic situation. This sentiment may influence financial markets and consumer behavior, and may be important for the forecasting models of policymakers like the Italian Treasury Department. As a result, economic sentiment can be used as a nearly instantaneous proxy to anticipate traditional economic indicators and exogenous variables. Previous studies in economic news sentiment have applied two distinct approaches: (1) evaluating the sentiment of the news content or (2) analyzing the tone with which the news is presented. Although each approach yields valuable perspectives in different settings, they have never been used in combination. Additionally, they usually rely on sentiment lexicons, which are predominantly designed for English, making them highly language-dependent and often insufficient for capturing the subtle nuances present in different economic subtopics. To address these limitations, we propose to combine the content sentiment in central banks communications and the tone sentiment in news articles. We train and use deep-learning classifiers tailored for sentiment analysis in the Italian language using a five-grade rating sentiment system. This approach enhances the models robustness and promotes seamless adaptation to different topics within the economic news domain.

H0201: **Forecasting Italian firms’ default probability using Prophet: A cycle-informed approach**
Presenter: **Eugenio Cangiano**, Sogei, Italy
Co-authors: Andrea Rollin, Daria Scacciatelli

The Italian Ministry of Economy and Finance plays a crucial role in monitoring public loan guarantee programs, which serve as key financial interventions to support enterprise liquidity and capitalisation. A critical aspect of this activity is the estimation and monitoring of default probabilities, essential for evaluating expected losses in guarantee portfolios and ensuring the efficient allocation of public resources. Default probabilities are estimated conditional on the economic cycle, using models that incorporate sectoral and macroeconomic variables to provide a forward-looking assessment of credit risk. Traditional econometric models often struggle to capture major economic shocks, including the pandemic crisis and financial downturns, due to their limited ability to handle structural breaks and nonlinear trends or dynamics. In contrast, the Prophet model offers a more flexible approach, effectively managing macroeconomic shocks and regime shifts. The analysis compares Prophet and traditional models across the most representative industrial sectors of a major Italian guarantee fund, evaluating the model’s contribution to improving default probability projections. The results highlight the advantages of integrating machine learning-based time series models into the econometric toolkit of financial institutions to enhance credit risk analysis and the strategic management of the public guarantees framework.

H0210: **ITFIN: A stock-flow consistent model for the Italian economy**
Presenter: **Valeria Macauda**, Sogei spa, Italy
Co-authors: Cristian Tegami

ITFIN is a quarterly, stock-flow consistent econometric model for the Italian economy developed at Italy's Department of Treasury. The model has two distinctive features: a strong focus on the financial and banking system and an emphasis on the macroeconomic consequences of sovereign risk. We model the financial position of each institutional sector in the economy with a stock-flow consistent approach to their balance sheets. In modelling both the supply and demand for many financial instruments, their prices and rates of return are derived along with a characterization of how financial stocks impinge on agents' decisions and the pattern of non-financial variables in the economy. This paper describes the features of ITFIN by illustrating the model structure and its main equations and identities. We also describe the properties of the model by simulating the economy's response under two counterfactual scenarios on the ECB's asset purchase programmes and after three different fiscal policy shocks.

H0200: **The macroeconomic impact of structural reforms: The case of Italy**
Presenter: **Sara DAndrea**, Sogei spa, Italy
Co-authors: Silvia DAndrea, Giovanni Di Bartolomeo, Paolo DImperio, Giancarlo Infantino, Mara Meacci

A methodology is proposed to map structural reforms from granular data to an aggregate model, exploring their transmission mechanisms and their macroeconomic and social impacts. The focus is on the rich case of the reforms associated with the Italian Recovery and Resilience Plan. We document a significant potential impact on medium- and long-term GDP and find that the labour market and education measures are the main drivers of the impact on GDP and employment. We also examine the distributional impact of the reforms on the functional income distribution.

HO018 **Room Lecture room 4** RECENT ADVANCES IN COMPLEX DATA ANALYSIS **Chair: Bojana Milosevic**

H0209: **Conformal changepoint localization**
Presenter: **Aaditya Ramdas**, Carnegie Mellon University, United States

Offline changepoint localization is the problem of estimating the index at which a change occurred in the data-generating distribution of an ordered list of data. We present the broadly applicable CONCH (CONformal CHangepoint localization) algorithm, which uses a matrix of conformal p-values to produce a confidence interval for a changepoint under the mild assumption that the pre-change and post-change distributions are each exchangeable. We exemplify the CONCH algorithm on a variety of synthetic and real-world datasets, including using black-box classifiers to detect changes in sequences of images or text.

H0211: **Efficient offline nonparametric changepoint detection for univariate data**
Presenter: **Dean Bodenham**, Imperial College London, United Kingdom

A novel offline changepoint detection method is introduced for identifying multiple changepoints in a univariate time series. This approach uses a popular nonparametric two-sample test combined with data structures from computer science to achieve $O(n\log n)$ computational complexity for identifying a single change. Current approaches to this problem tend to be either parametric, requiring distributional assumptions about the data, or nonparametric with a higher computational cost. Empirical results are presented comparing our approach to well-known changepoint detection methods in a variety of scenarios.

H0202: **Energy-distance-based two-sample testing in the presence of incomplete data**
Presenter: **Bojana Milosevic**, University of Belgrade, Serbia
Co-authors: Danijel Aleksic

The problem of two-sample testing is addressed in the presence of missing data under general missingness mechanisms. The focus is placed on the widely used energy-based two-sample test. In addition to the standard complete case analysis, we introduce a novel adaptation of the test statistic that incorporates all available data, as well as two resampling procedures for p-value approximation. Furthermore, we present a new bootstrap procedure tailored for scenarios where the test statistic is computed on imputed data using standard imputation techniques. Through a comprehensive simulation study, the proposed methods are evaluated across a range of sample sizes, dimensions, data distributions, missingness mechanisms, and missing data proportions. Practical guidelines are offered based on the observed performance in each scenario.

H0199: **Parking on random spanning tree in a Hilbert space with applications to tree-based machine learning**
Presenter: **Andrej Srakar**, Institute for Economic Research Ljubljana, Slovenia

Parking problems on trees have found interesting probability applications in recent years. Yet, many research questions still lack adequate research. The aim is to study the parking problem for a random spanning tree, spreading in a Hilbert space. We will develop scaling limits, large deviations for such a random process, and appropriate limit theorems for the behaviour of parking components. We will point to possible extensions to frozen Erdos-Renyi random graph processes and more general forms of Galton-Watson trees. In application, we will translate and use the theoretical findings for modelling in machine learning using tree-based models, such as random forest or ensemble methods.

Tuesday 08.07.2025 14:00 - 15:40 Parallel Session C – HiTECCoDES2025

HO015 Room Amphitheater 1 MACHINE LEARNING FOR HIGH DIMENSION TIME SERIES Chair: Weining Wang

H0154: **Matrix-valued factor model with time-varying main effects**
Presenter: **Clifford Lam**, London School of Economics and Political Science, United Kingdom
Co-authors: Zetai Cen

The matrix-valued time-varying Main Effects Factor Model (MEFM) is introduced. MEFM is a generalisation of the traditional matrix-valued factor model (FM). We give rigorous definitions of MEFM and its identifications. We propose estimators for the time-varying grand mean, row and column main effects, and the row and column factor loading matrices for the common component. Rates of convergence for different estimators are spelt out, with asymptotic normality shown. The core rank estimator for the common component is also proposed, with the consistency of the estimators presented. We propose a test to test if FM is sufficient against the alternative that MEFM is necessary, and demonstrate the power of such a test in various simulation settings. We also demonstrate numerically the accuracy of our estimators in extended simulation experiments. A set of NYC Taxi traffic data is analysed, and our test suggests that MEFM is indeed necessary for analysing the data against a traditional FM.

H0164: **A modified VAR-deGARCH model for asynchronous multivariate financial time series via variational Bayesian inference**
Presenter: **Ray-Bing Chen**, National Tsing Hua University, Taiwan

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

H0189: **Arellano-Bond LASSO estimator for dynamic linear panel models**
Presenter: **Chen Huang**, Aarhus University, Denmark
Co-authors: Victor Chernozhukov, Ivan Fernandez-Val, Weining Wang

The Arellano-Bond estimator is a fundamental method for dynamic panel data models, widely used in practice. However, the estimator is severely biased when the data's time series dimension T is long due to the large degree of overidentification. We show that weak dependence along the panel's time series dimension naturally implies approximate sparsity of the most informative moment conditions, motivating the following approach to remove the bias: First, apply LASSO to the cross-section data at each time period to construct most informative (and cross-fitted) instruments, using lagged values of suitable covariates. This step relies on approximate sparsity to select the most informative instruments. Second, apply a linear instrumental variable estimator after first differencing the dynamic structural equation using the constructed instruments. Under weak time series dependence, we show the new estimator is consistent and asymptotically normal under much weaker conditions on T's growth than the Arellano-Bond estimator. Our theory covers models with high-dimensional covariates, including multiple lags of the dependent variable, which is common in modern applications. We illustrate our approach by applying it to weekly county-level panel data from the United States to study opening K-12 schools and other mitigation policies' short and long-term effects on COVID-19's spread.

H0212: **Testing for spurious factor analysis on high dimensional nonstationary time series**
Presenter: **Yi He**, University of Amsterdam, Netherlands
Co-authors: Bo Zhang

Spurious factor behaviors arise in large random matrices with high-rank random signal components and heavy-tailed spectral distributions. The aim is to establish analytical probabilistic limits and a distribution theory for these spurious behaviors in high-dimensional non-stationary time series. We transform scree plots into Hill plots to detect spectral patterns in these spurious factor models and develop max-t tests to distinguish between spurious and genuine factor models. Simulations confirm the excellent size and power performance of our test in finite samples. Applying the tests to three real-life datasets, we detected spurious factors in both economic and climate data, and genuine factors in finance data.

HO006 Room Amphitheater 2 COMPLEX RESEARCH DESIGNS Chair: Kalliopi Mylona

H0178: **Beyond the Pareto front: A TOPSIS-based framework for multi-criteria design with the multiDoE R Package**
Presenter: **Matteo Borrotti**, University of Milan-Bicocca, Italy

Optimizing a single criterion, such as D-optimality or A-optimality, is no longer sufficient in many experimental scenarios. Real-world applications often involve conflicting design criteria that must be jointly balanced. A novel application of the TOPSIS (Technique for Order Preference by Similarity to Ideal Solution) method is introduced within the multi-criteria design of experiments framework, as implemented in the CRAN-published R package multiDoE (https://cran.r-project.org/web/packages/multiDoE/index.html). After a brief overview of classical selection strategies on the Pareto front, we present the TOPSIS-based ranking approach regarding interpretability, stability, and performance. The methodology is illustrated through real and simulated case studies involving multiple optimality criteria. How the multiDoE package can facilitate reproducible workflows in multi-criteria DoE scenarios is also discussed.

H0186: **Bayesian design of experiments for unmanned aerial vehicles path planning**
Presenter: **Yiolanda Englezou**, University of Cyprus, Cyprus
Co-authors: Stelios Timotheou, Christos Panayiotou

Traffic monitoring is one of the major tools used for transportation operations and planning. With the emergence of Unmanned Aerial Vehicles (UAVs), new capabilities for enhancing traffic management have emerged. Despite their potential, UAV applications in traffic management have primarily focused on sporadic surveillance of road networks and historical traffic data extraction. Path planning stands out as a critical challenge for UAVs, aiming to optimise routes from initial to target points. For such complex tasks, an optimal design often depends on uncertain model parameters. This dependence leads naturally to a Bayesian approach which can (a) make use of any prior information, and (b) be tailored to the reduction of posterior uncertainty. Towards this, we propose an online Bayesian optimal UAV trajectory design methodology. The proposed method strategically selects the next UAV sampling points to obtain traffic density measurements, while minimising the total uncertainty of the traffic density across all time-space points within the studied time-horizon. The proposed approach integrates the Gaussian Process model into a Bayesian framework to accurately estimate traffic density in multi-lane highways, considering both temporal and spatial correlations, even when data points are sparse. Employing a decision-theoretic approach, we develop a Bayesian optimal UAV trajectory design scheme to mitigate uncertainty.

H0214: **Optimized recovery sampling to test for missing not at random**
Presenter: **Robin Mitra**, UCL, United Kingdom

Missing data can lead to inefficiencies and biases in analyses, in particular when data are missing not at random (MNAR). It is thus vital to understand and correctly identify the missing data mechanism. Recovering missing values through a follow-up sample allows researchers to conduct hypothesis tests for MNAR, which is not possible when using only the original incomplete data. Investigating how the properties of these tests are affected by the follow-up sample design has not been explored in the literature. The results provide comprehensive insight into the properties of one such test based on the commonly used selection model framework. Conditions are determined for recovery samples that allow the test to be applied appropriately and effectively, i.e., with known type I error rates and optimized with respect to power. An integrated framework is thus provided for testing the presence of MNAR and designing follow-up samples in an efficient, cost-effective way. The methodology's performance is evaluated through simulation studies and on a real data sample.

H0185: **NeuroBayes Design Optimizer (NBDO) for high-dimensional settings**

Presenter: **Theodoros Ladas**, King's College London, United Kingdom

Co-authors: Caterina May, Kalliopi Mylona, Davide Pigoli

Finding an optimal experimental design is computationally challenging, especially in high-dimensional spaces. To tackle this, we introduce the NeuroBayes Design Optimizer (NBDO), which uses neural networks to find optimal designs for high-dimensional models, by reducing the dimensionality of the search space. This approach significantly decreases the computational time needed to find a highly efficient optimal design, as demonstrated in various numerical examples. Comparisons with the Coordinate Exchange (CE) algorithm are presented. The method offers a balance between computational speed and efficiency, laying the groundwork for more reliable design processes.

HO011 Room Lecture room 4 ADVANCES IN STATISTICS AND FINANCIAL ECONOMETRICS Chair: Sandra Paterlini

H0181: **The role of biodiversity risk in shaping bank lending decisions**

Presenter: **Karoline Bax**, Technical University of Munich, Germany

Co-authors: Aida Cehajic

The aim is to examine the extent to which banks incorporate firm-level biodiversity risk into their lending decisions, both in terms of loan pricing and loan volume adjustments. We use a comprehensive syndicated loan market dataset that captures lender-borrower relationships and allows for matching with firm-level biodiversity risk measures. Beyond employing existing biodiversity risk measures, we extend the methodology by developing a more granular text-based biodiversity risk measure. This novel metric is constructed using cosine similarity of textual embeddings, allowing for a nuanced assessment of biodiversity-related corporate disclosures. Furthermore, we explore potential transmission mechanisms through which biodiversity risk is reflected in lending decisions, analyzing whether banks explicitly price biodiversity risk or adjust credit supply based on borrowers' exposure to biodiversity-related vulnerabilities. We examine whether bank location, UNEP-FI membership, and borrower characteristics, such as low ESG scores and environmental violations, affect lending behavior. While we find clear evidence that banks penalize biodiversity risk through higher spreads and lower loan volumes, the underlying channels remain inconclusive.

H0218: **Improving index tracking and portfolio performance with regularization**

Presenter: **Dietmar Maringer**, University of Basel, Switzerland

Co-authors: Sandra Paterlini

Historical data are commonly used for portfolio optimization when parametric models fail to capture essential features of asset returns. However, this approach poses a challenge: long time series may include outdated observations that no longer reflect current market conditions, while short series with recent data suffer from small-sample issues and risk overfitting, especially when the asset universe is large. Recent research shows that regularization techniques, such as Lasso (L1), Ridge (L2), or Elastic Net (combining L1 and L2), can enhance portfolio construction. Yet, practical guidance on calibration remains limited, and applications to passive management strategies are rarely explored. This paper addresses these gaps by investigating different regularization techniques in the context of index tracking and enhanced indexing. We conduct empirical studies to evaluate how different calibration choices affect portfolio performance, considering various objectives, training window lengths, and out-of-sample investment horizons. Our findings indicate that regularization can improve out-of-sample tracking error and related metrics, though the magnitude of improvement depends heavily on calibration choices and additional optimization objectives beyond tracking error.

H0220: **Recovering network hubs with PCGLASSO: Theory, algorithm, and performance**

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

Co-authors: Adam Chojecki, Ivan Hejny, Bartosz Kolodziejek, Jonas Wallin

Regularization techniques have become standard tools for estimating high-dimensional precision matrices and graphical models. These methods typically penalize the magnitudes of the elements in the precision matrix, which leads to a lack of scale invariance. This means that the structure of the estimated graphical model may change depending on how the variables are scaled. To overcome this limitation, the Partial Correlation LASSO (PCGLASSO) has been recently proposed. This method applies regularization directly to the elements of the partial correlation matrix. We will introduce a novel algorithm for PCGLASSO that is substantially more efficient than existing approaches. We will also present new theoretical results addressing convexity and consistency, including Irrepresentability Conditions for accurate graph structure recovery. Our theoretical insights, supported by empirical evidence, show that PCGLASSO significantly outperforms GLASSO, particularly in identifying key hub structures within the network.

H0155: **An l0-constrained and l1-regularized estimator for graphical models**

Presenter: **Sandra Paterlini**, University of Trento, Italy

Co-authors: Alessandro Fulci, Emanuele Taufer

Graphical models provide a versatile framework for representing conditional dependencies among random variables, with the precision matrix playing a central role in capturing these relationships. Traditional estimation methods, such as the Graphical Lasso (Glasso) with l1 regularization, suffer from key limitations, including significant bias, absence of grouping effects, and sensitivity to regularization parameters. On the other hand, algorithms leveraging l0 regularization can address some of these drawbacks but often lack shrinkage, leading to potential instability. To overcome these challenges, we introduce the Sparsity Constrained Graphical Lasso (SCGlasso), a novel estimator that integrates an l0 constraint on the number of non-zero elements with an l1 penalty. This hybrid approach decouples sparsity promotion from shrinkage, effectively addressing the limitations of l1 and l0 regularization in isolation. We design a coordinate descent algorithm for SCGlasso and establish its convergence to a local minimum. Through extensive simulations, we benchmark SCGlasso against established methods, including Glasso, Selo, and Atan penalties. The results highlight its superior performance in model selection accuracy and robustness, particularly in small sample scenarios. Finally, we demonstrate its practical utility through an application to gene expression data, showcasing the method's effectiveness in uncovering meaningful conditional dependency structures.

Tuesday 08.07.2025 16:10 - 17:25 Parallel Session D – HiTECCoDES2025

HO009 Room Amphitheater 1 TOPICS IN FINANCIAL ECONOMETRICS Chair: Leopold Soegner

H0153: **On the influence of the choice of seasonal adjustment method on forecasting national accounts aggregates across the EU**
Presenter: **Martin Ertl**, Institute for Advanced Studies, Austria
Co-authors: Robert Kunst, Adrian Wende

Empirical macroeconomic research routinely relies on seasonally adjusted data. For the most part, two methods of seasonal adjustment are used today: The moving average X-11 method (and its offsprings, such as X-12 and X-13) and the SEATS method that is based on ARIMA modeling. We study which of the two classes of methods assists in obtaining more accurate forecasts of annual targets, and in which circumstances it is better not to seasonally adjust the data at all. We investigate this question empirically and with Monte Carlo simulations, and with both data-driven ARIMA models and theoretically grounded dynamic stochastic general equilibrium models.

H0166: **Online breakpoint-detection in cointegrating relationships**
Presenter: **Leopold Soegner**, Institute for Advanced Studies, Austria
Co-authors: Martin Wagner

A closed-end consistent monitoring procedure is developed with the goal of detecting structural changes in cointegrating relationships. We consider a vector error correction model and allow for different specifications of the deterministic terms. We consider $\beta' y_t$, where β is a matrix containing the cointegrating vectors and (y_t) is a process integrated of order one. We propose a monitoring test statistic to investigate the stability of these cointegrating relationships. We obtain the asymptotic distribution of our test statistic under the null hypothesis of no structural breaks. A calibration period is used for parameter estimation, after which online break-point detection is performed. The procedure stops at the first time point at which the test statistic exceeds the corresponding critical value. A simulation study is provided to investigate the finite sample properties of our monitoring procedure.

H0192: **Monitoring structural breaks in vector autoregressive models**
Presenter: **Masoud Abdollahi**, Institute for Advanced Studies, Austria

A closed-end monitoring approach is proposed for the online detection of structural breaks in non-stationary cases. The framework is based on an error correction model that accommodates potential breaks in both the cointegrating relationships and adjustment vectors, with the possibility of a changing cointegration rank. To this end, we develop Lagrange-multiplier statistics tailored for sequential monitoring. Following a calibration period used to estimate the model parameters, monitoring begins and terminates once the test statistic crosses its corresponding critical value. Through an extensive simulation study, we investigate the performance of the monitoring procedure, including assessments of the observed size and power of the test.

HO019 Room Amphitheater 2 BAYESIAN METHODS, DECISION-MAKING, AND EXPERIMENTAL DESIGNS Chair: Matteo Borrotti

H0172: **Bayesian optimization on the space of symmetric positive definite matrices**
Presenter: **Federico Pavesi**, University of Milan, Bicocca, Italy
Co-authors: Antonio Candelieri

An extension of Bayesian Optimization for functionals over symmetric positive definite matrices is presented. The proposed method leverages the log-euclidean product, which endows the manifold with a Lie group structure and a bi-invariant metric. This allows the construction of a metric-based positive definite kernel, which is a naive generalization of the well-known squared exponential kernel. Although the new kernel is non-stationary and generally unrelated to the squared exponential kernel in the Euclidean space, its simple form makes it attractive for Bayesian Optimization in non-Euclidean spaces, since its gradient can be computed in closed form. A Gaussian Process based on the new kernel approximates the functional of interest while the acquisition function is optimized via gradient-based methods. Both Expected Improvement and Lower Confidence Bound are considered in the paper. Experiments consider the minimization of simple functionals, like the Wasserstein distance, with results empirically proving that the proposed method is able to identify minima, even if the non-stationarity of the kernel negatively affects the quality of function approximation. Future research may focus on improving kernel design to enhance approximation accuracy while preserving optimization efficiency.

H0176: **Conformal prediction for safe decision-making**
Presenter: **Valentina Zangirolami**, University of Milano-Bicocca, Italy
Co-authors: Matteo Borrotti, Antonio Candelieri

Safety is the core feature to avoid system disruptions while aiming to develop more efficient policies that improve existing human-expert control strategies. Safe conformal constraints for decision-making in dynamic systems, leveraging split conformal prediction to provide robust uncertainty quantification for next-state predictions. By making no distributional assumptions, Conformal Prediction ensures coverage guarantees for safe constraints, formulated as no next-state violations. We employ normalized nonconformity measures to integrate conformal prediction intervals in the decision rules to obtain point-wise prediction intervals. We extend the framework with Mondrian Conformal Prediction techniques to fully adapt conformal intervals in dynamic systems while maintaining validity. The efficacy of this approach is then empirically demonstrated on an optimal control task for a water tank system, comparing the proposed safe (conformal) policy against both a baseline safe policy and a state-of-the-art alternative lacking conformal predictions. Results highlight the effectiveness of our method in balancing safety and performance.

H0182: **Optimal experimental designs for function-on-function linear models**
Presenter: **Kalliopi Mylona**, King’s College London, United Kingdom
Co-authors: Caterina May, Theodoros Ladas, Davide Pigoli

The function-on-function linear model has broad applications in science, industry, and technology. Examples can be found in the pharmaceutical sector, the environment, engineering, chemistry, etc. We propose a Bayesian optimality criterion for constructing optimal experimental designs where both the response and the factors have a functional nature. Several examples of optimal designs are provided and discussed.

HO012 Room Lecture room 4 HIGH-DIMENSIONAL PROBABILITY AND MACHINE LEARNING Chair: Andrej Srakar

H0213: **Spectral estimators for multi-index models**
Presenter: **Yihan Zhang**, University of Bristol, United Kingdom

Multi-index models provide a popular framework to investigate the learnability of functions with low-dimensional structure. Due to their connections with neural networks, they have been an object of recent intensive study. The focus is on recovering the subspace spanned by the signals via spectral estimators – a family of methods that are routinely used in practice, often as a warm-start for iterative algorithms. The main technical contribution is a precise asymptotic characterization of the performance of spectral methods, when sample size and input dimension grow proportionally and the dimension p of the space to recover is fixed. Specifically, we locate the top-p eigenvalues of the spectral matrix and establish the overlaps between the corresponding eigenvectors (which give the spectral estimators) and a basis of the signal subspace. Our analysis unveils a phase transition phenomenon in which, as the sample complexity grows, eigenvalues escape from the bulk of the spectrum and, when that happens, eigenvectors recover directions of the desired subspace. The precise characterization we put forward enables the optimization of the data preprocessing, thus allowing us to identify the spectral estimator that requires the minimal sample size for weak recovery.

H0216: **Gradient span algorithms make predictable progress in high dimension**

Presenter: **Felix Benning**, University of Mannheim, Germany

Co-authors: Leif Doering

The aim is to prove that all gradient span algorithms have asymptotically deterministic behavior on Gaussian random objective functions as the dimension tends to infinity. This result explains the counterintuitive phenomenon that different training runs of many large machine learning models result in approximately equal optimization-progress curves despite random initialization on a complicated non-convex landscape.

H0222: **Partial dependence and functional principal component-based reconstruction for explainable functional random forests**

Presenter: **Fabrizio Maturo**, Universita Telem.universitas Mercatorum, Italy

Co-authors: Annamaria Porreca

Functional random forests (FRF) combine the strengths of ensemble learning with functional data analysis, offering strong predictive performance on high-dimensional functional datasets. However, their limited transparency poses a major barrier in critical applications. This contribution introduces a set of explainability tools that extend classical partial dependence plots to the functional context, through functional partial dependence plots (FPDPs), enabling the study of the marginal effect of each functional principal component (FPC) score. To support interpretation, FPDPs are paired with graphical reconstructions of the functional shape associated with score variations. The approach highlights how individual FPCs influence predictions both in score space and in the time domain. Further, model-specific and model-agnostic FPC importance measures are provided, including an integrated visual tool comparing internal and external relevance. Applied to ECG signals, the method reveals meaningful patterns that support the interpretation of model behavior. These tools help bridge the gap between accuracy and explainability in functional machine learning.

Wednesday 09.07.2025 09:05 - 10:20 Parallel Session F – HiTECCoDES2025

HI023 Room Amphitheater 1 HITEC SESSION: BAYESIAN METHODS AND TEXT MINING Chair: Bernardo Nipoti

H0183: **Evolution of prevalence and dominance of HiTEc topics**
Presenter: **Louisa Kontoghiorghes**, Kings College London, United Kingdom
Co-authors: Ana Colubi, George Kapetanios

Text analysis is used to track the evolution of themes within the COST Action HiTEc on a scientific conference's Book of Abstracts (BoAs). To represent HiTEc's relevant themes, a set of keywords is automatically extracted from its proposal. A new topic modeling method is used, the time-varying weighted Latent Dirichlet Allocation (tvwLDA), which estimates the term distribution of the topics and topic distribution of the documents at each time index, enabling the tracking of the topic evolution of the BoAs over time. After applying tvwLDA, the extracted keywords are used to estimate topic prevalence and dominance of HiTEc's themes. The prevalence measures the frequency of HiTEc's themes, while the dominance, a new estimator, combines the topic prevalence with the Simpson index to capture the abundance of HiTEc's related themes in the BoAs.

H0191: **Sampling uncertainty of research topics**
Presenter: **Anna Staszewska-Bystrova**, University of Lodz, Poland
Co-authors: Viktoriia Naboka-Krell, Victor Bystrov, Peter Winker

In latent topic models, estimated topic-word and document-topic probabilities are typically reported with no indication of sampling uncertainty. The lack of additional information on sampling uncertainty might result in misleading conclusions regarding topic structure and prevalence. We propose to measure sampling uncertainty using a bootstrap method and describe how uncertainty can be captured by novel types of word clouds reporting topic-word probability estimates and by confidence bands designed for reporting time series estimates of topic weights. The application of the new measures and methods is illustrated with an empirical example involving conference abstracts. The results indicate varying robustness of estimated research topics with respect to resampling of documents from the same text collection. In particular, some estimated topics may not persist across resampled corpora, and the estimation precision of topic-word probabilities within the same topic can exhibit significant variation. Similar uncertainty is associated with topic prevalence over time. The proposed confidence bands for dynamic topic weights can be used to make inferences about structural changes in research topic trends.

H0228: **Bayesian modeling of multiple network data**
Presenter: **Bernardo Nipoti**, University of Milan Bicocca, Italy
Co-authors: Francesco Barile, Simon Lunagomez

A flexible framework is discussed for modeling multiple network data using similarity metrics to compare networks. Within this setting, we introduce a novel Bayesian nonparametric model that identifies clusters of networks with similar connectivity patterns. Our approach is based on a location-scale Dirichlet process mixture of centered Erdos-Renyi kernels, where each component is defined by a representative network, or mode, and a univariate measure of dispersion around it. This model offers desirable properties, including full support in the Kullback-Leibler sense and strong consistency. For posterior inference and network clustering, we develop an efficient Markov chain Monte Carlo algorithm. The models performance is evaluated through extensive simulations and applications to human brain network data and a global food trade dataset.

HO021 Room Amphitheater 2 ADVANCED STATISTICAL TOOLS IN RISK MANAGEMENT Chair: Massimiliano Caporin

H0159: **Optimizing portfolios through ESG risk budgeting**
Presenter: **Massimiliano Caporin**, University of Padova, Italy
Co-authors: Monica Billio, Sandra Paterlini, Runfeng Yang

A novel ESG risk budgeting framework is introduced for constructing ESG-optimal portfolios, focusing on managing ESG risk contributions rather than pursuing ESG performance. This framework is tailored for investors prioritizing ESG risk management over non-pecuniary ESG goals. We apply this framework to the European stock market from 2013 to 2022. We first compare optimization outcomes across various ESG risk targets and find significant risk-return trade-offs when actively managing ESG risk exposure. Additionally, we observe that the choice of ESG data provider influences the risk budgeting results under active risk management. Then, we apply the framework to evaluate the impact of ESG risk on the stock market. We find that the impact of ESG risk can be large in certain sector groups and during extreme ESG-related market events, though the impact remains limited. Our findings highlight the complexities involved in incorporating ESG factors into investment strategies.

H0160: **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 for multi-asset allocation is studied 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 in this study 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 risk structure changes and maintain a balanced exposure to the then prevailing uncorrelated risk sources. This tool can help a portfolio manager to understand and choose those risk sources that have earned risk, focusing on those risk factors.

H0173: **The role of local and global economic policy uncertainty in GARCH-MIDAS forecasts of US state-level volatility**
Presenter: **Vincenzo Candila**, University of Salerno, Italy
Co-authors: Oguzhan Cepni, Giampiero Gallo, Rangan Gupta

The aim is to examine the influence of local (state-specific) and global Economic Policy Uncertainty (EPU) on the volatility of US state-level equity returns. We employ a GARCH-MIDAS approach incorporating multiple EPU indices as low-frequency predictors of daily stock return volatility. To address the challenge of selecting the most relevant EPU indices, we utilize an Elastic Net (EN) shrinkage method to integrate forecasts from different models. The results reveal that the proposed model, which leverages information from both local and global EPU indices, generally outperforms single specifications. Further, a cluster analysis based on the volatility forecasts uncovers distinct geographical patterns, suggesting that state-level volatility is influenced by both state-specific and nationwide policy uncertainties. These findings highlight the importance of considering both local and global EPU in understanding and predicting the volatility dynamics at the regional level.

Wednesday 09.07.2025 10:50 - 12:30 Parallel Session G – HiTECCoDES2025

HI046 Room Amphitheater 1 HITEC SESSION: ECONOMETRICS AND MACHINE LEARNING Chair: Christina Erlwein-Sayer

H0157: **Enhancing economic forecasts through Hilbert space projection methods**
Presenter: **Tsvetomira Tsenova**, Experian Bulgaria, Bulgaria

Improving forecast accuracy of economic and financial variables is essential in central banking and financial institutions, as conditioning decisions on economic developments has become obligatory. Forecast reports contain the entire probability distribution of the target variables, i.e., point, uncertainty, risk, derived probabilistic scenarios, at near, medium, and long-term horizons. However, traditional forecasting methods are known to converge to their equilibrium much too quickly. While having sensible equilibrium properties is desirable in banking and finance, the decision-makers are deprived of the opportunity to design an optimal reaction to the shorter-term predictable volatility. The aim is to enhance the applied forecasting process by exploiting the Hilbert Space Projection methods. Advantages and disadvantages of various forecasting methods are discussed from a practical perspective, and optimal ways to combine them.

H0215: **Advancing Markowitz: Asset allocation forest**
Presenter: **Alla Petukhina**, ASE Bucharest, Romania
Co-authors: Anastasija Tetereva

A novel Asset Allocation Forest (AAF) model is proposed that combines the well-established machine learning (ML) tool with the conventional portfolio optimization method. The determination of locally optimal portfolio weights, which dynamically respond to market conditions, effectively captures market regimes, structural breaks, and smooth transitions in a data-driven manner. We illustrate the proposed model using a multi-asset portfolio consisting of equities, bonds, credit, high yield, and commodities. The AAF consistently outperforms established benchmarks, including the Hidden Markov Model (HMM), even when trading costs are taken into account. It also opens the door to valuable economic insights. By constructing accumulated local effects (ALE) plots, we find evidence of flight-to-safety, suggesting a strategic shift from riskier assets to less volatile bonds during periods of increased market turbulence. Furthermore, our model shows a pronounced preference for bonds in inflationary periods, demonstrating its adaptability to different economic conditions.

H0223: **Statistical methods and supervised-unsupervised machine learning for EEG signal analysis and detection of brain injury**
Presenter: **Robertas Alzbutas**, Kaunas University of Technology, Lithuanian Energy Institute, Lithuania
Co-authors: Vaida Abraskeviciute, Asta Kybartaite-Ziliene, Giedre Alzbutiene

The aim is to present an integrated approach to neonatal EEG signal analysis for the detection of hypoxic-ischemic brain injury, combining statistical methods with supervised and unsupervised machine learning. The first part of the work emphasizes predictive modeling based on preprocessed EEG data, employing classification algorithms like support vector machines and neural networks. Statistical signal processing techniques were applied to extract relevant features, enabling automated identification of abnormal patterns and supporting early-stage clinical decision-making. These supervised models demonstrate potential to complement expert assessments and improve the timeliness and consistency of diagnosis. In the second part, unsupervised learning methods were applied to EEG data from neonates diagnosed with hypoxic-ischemic encephalopathy. After artifact removal and dimensionality reduction, clustering algorithms explored latent structure in the data, enabling grouping based on EEG signal characteristics without labeling. While detailed grading posed challenges, a simplified binary clustering strategy yielded clinically relevant groupings comparable to supervised classification outcomes. The results highlight the promise of combining data-driven unsupervised methods with domain knowledge in medical diagnostics. Overall, this dual approach demonstrates the value of integrating statistical modeling and machine learning to enhance EEG-based neurodiagnostic tools in neonatal care.

H0225: **Explaining switching behavior: Consumer attention and choice in car insurance market**
Presenter: **Kadri Maennasoo**, Tallinn University of Technology, Estonia
Co-authors: Kaido Kepp

The focus is on the vehicle lessees' Motor Own Damage insurance search frictions and choices. We use a consumer-level annual panel of policy and insurance offers data over 2010-2018 from the biggest insurance broker, consolidating the Estonian car insurance market offers. We apply the two-stage discrete choice model, building on previous work that identifies the sources of consumer inertia by separating the attention and choice decisions given the observed switches to new providers. Consumers choose from a set of pre-listed offers corresponding to the best alternative choices available in the market. Our results show strong inertia that stems from consumer inattention and considerable heterogeneity of inattention across consumer groups. Consumers' decisions to switch or stay with the current provider reveal substantial price elasticity and only a modest effect of the insurance providers' brand preference. Multiple robustness checks confirm adequate prediction and plausibility of our estimates.

HO010 Room Amphitheater 2 STATISTICAL METHODS FOR COMPLEX DATA STRUCTURES Chair: Maria Brigida Ferraro

H0167: **Clustering locally stationary time series using quantile autocorrelations**
Presenter: **Angel Lopez Oriona**, King Abdullah University of Science and Technology (KAUST), Saudi Arabia
Co-authors: Ying Sun, Jose Vilar

Locally stationary time series frequently arise in various fields, such as environmental sciences, economics, and seismology. However, statistical methods for analyzing locally stationary time series remain underdeveloped. A clustering approach is presented for local stationary time series that uses a dissimilarity measure based on local estimates of the quantile autocorrelation function at each time point. This distance is then combined with a K-medoids-type minimization problem, which incorporates a penalty term driven by the neighborhood size considered for the local estimation. To solve this problem, a three-step iterative procedure that guarantees a decrease in the objective function at each iteration is proposed. Several simulations show that the method generally outperforms some natural benchmarks in terms of clustering accuracy. The potential of the approach is demonstrated through an interesting application involving real-time series.

H0177: **Clustering models for multi-view data**
Presenter: **Paolo Giordani**, Sapienza University of Rome, Italy

It is common to encounter situations where variables are observed on a set of objects from multiple sources. In such cases, data are usually referred to as multi-view. Multi-view data offer a remarkably rich source of information, provided that suitable models are employed, as standard techniques often fall short. While models for multi-view data are frequently used for dimensionality reduction through components, clustering has garnered considerable attention in recent years. Typically, this involves clustering objects and compressing variables and sources through components. The clustering step is often performed using a hard partitioning approach, where objects are assigned exclusively to a single cluster. However, this can yield counterintuitive outcomes, particularly when objects exhibit characteristics shared across multiple clusters, which is frequently the case. Therefore, it is advantageous to partition objects using a fuzzy approach, allowing for soft membership degrees. After reviewing existing clustering models for multi-view data, fuzzy variants will be presented and examined through case studies.

H0180: **funBIalign: A hierarchical algorithm for functional motif discovery**
Presenter: **Marzia Cremona**, Universite Laval, Canada
Co-authors: Jacopo Di Iorio, Francesca Chiaromonte

Motif discovery is gaining increasing attention in the domain of functional data analysis. Functional motifs are typical shapes or patterns that recur multiple times in different portions of a single curve and/or in misaligned portions of multiple curves. We define functional motifs using an additive model and propose funBIalign for their discovery and evaluation. Inspired by clustering and biclustering techniques, funBIalign is a multi-step procedure that uses agglomerative hierarchical clustering with complete linkage and a functional distance based on mean squared residue scores to discover functional motifs, both in a single curve (e.g., time series) and in a set of curves. We assess its performance and compare it to other recent methods through extensive simulations. Moreover, we use funBIalign for discovering motifs in two real-data case studies; one on food price inflation and one on temperature changes. Finally, we introduce another definition of motifs based on a multiplicative model that includes the more challenging scenario of motifs composed of portions sharing the same shape but having different amplitudes, and we extend funBIalign to discover amplitude-invariant functional motifs.

H0196: **Estimation of a simple linear regression model for random star-shaped sets**
Presenter: **Jose Grana Colubi**, University of Oviedo, Spain
Co-authors: Gil Gonzalez-Rodriguez, Ana Belen Ramos-Guajardo

The estimation of a simple linear regression model is undertaken when both the independent and dependent variables are star-shaped set-valued random elements. The suggested regression model is defined by using the set arithmetic, assuming that the components representing location and imprecision of the random elements in the model are handled independently. Once the theoretical framework is established, the least squares estimation for the linear model is performed, taking into account an appropriate distance within the space of star-shaped sets. This approach results in a constrained minimization problem, which is analytically solved. Furthermore, the strong consistency of the obtained estimators is analyzed, and a simulation study is carried out.

HO014 **Room Lecture room 4** **RECENT DEVELOPMENTS IN HYPOTHESIS TESTING** **Chair: Marija Cuparic**

H0193: **Comparision of k regression curves in R using a new R package**
Presenter: **Nora Villanueva**, University of Vigo, Spain

An R package is presented to test the hypothesis of equality of the k regression functions. The test is based on the comparison of two estimators of the distribution of the errors in each population. Kolmogorov-Smirnov and Cramr-von Mises type statistics are considered. A bootstrap procedure is used to approximate the critical values of the test. The R package is designed along lines similar to those of other R packages. Numerical summaries of the fitted objects can be obtained by using the print or summary functions. The main function of the package is equalreg(), which enables users to test the null hypothesis of equality of the k curves. In view of the high computational cost entailed in the bootstrap-based testing procedure, parallelization techniques are included to make it feasible and efficient in real situations. In addition, graphical outputs can be displayed based on ggplot2 package. Particularly, the autoplot() function lets the user draw the resulting estimated curves and the empirical cumulative distribution functions for the residuals of each group.

H0203: **Independence testing for mixed-type data using distributional transformations**
Presenter: **Marija Cuparic**, University of Belgrade, Serbia
Co-authors: Dana Bucalo Jelic, Bojana Milosevic

The problem of testing independence is addressed in mixed-type data settings where some components are discrete and others are positive and absolutely continuous. Two testing procedures are developed based on the recently introduced Barnighausen-Gaigal transformation, which characterises the joint distribution of such data. The asymptotic properties of the proposed tests are derived, and their practical performance is assessed through an extensive power study, which demonstrates the competitiveness and flexibility of the new approach.

H0206: **On resampling tests and the nested simulation problem**
Presenter: **Daniel Gaigall**, FH Aachen University of Applied Sciences, Germany
Co-authors: Julian Gerstenberg

Statistical tests based on resampling procedures are considered. A general framework is introduced that covers, in particular, bootstrap and permutation techniques for the computation of approximate quantiles as critical values in model specification testing. For the investigation of properties of such tests, Monte-Carlo simulation studies are customary. The resampling procedure leads to a nested simulation and ultimately to a nested simulation estimator for the rejection probability of the test. Choosing both the number of replications and the size of the simulation study large results in a considerable computational effort. To circumvent this problem, the so-called warp-speed method has become popular recently. For that reason, the related warp-speed estimator is revisited. Besides, the latest results for the nested simulation estimator indicate that a moderate or even rather small number of replications is sufficient to obtain useful simulation results. This enables a substantial reduction of the computational effort.

H0208: **A new class of goodness-of-fit tests using weighted degenerate U-statistics**
Presenter: **Marko Obradovic**, University of Belgrade, Serbia
Co-authors: Katarina Halaj, Bojana Milosevic

A novel flexible class of goodness-of-fit tests is introduced for univariate and multivariate data based on weighted degenerate U-statistics. The test statistics are constructed as linear combinations of U-statistics, with weights chosen to standardize the contribution of each component. We derive their asymptotic distributions under the null hypothesis and evaluate their general performance through simulations. As concrete examples, we demonstrate the method's application to testing multivariate normality and von Mises distribution on the circle, showing competitive power against existing alternatives while maintaining proper control of empirical size.

Wednesday 09.07.2025 14:00 - 15:40 Parallel Session H – HiTECCoDES2025

HI045 Room Amphitheater 1 HITEC SESSION: NONPARAMETRIC AND FUNCTIONAL-BASED METHODS Chair: Matus Maciak

H0224: **Combining nonparametric functionals for more effective decision-making and inference**
Presenter: **Arne Bathke**, University of Salzburg, Austria

Nonparametric statistical methods are usually characterized by rather generous invariance properties, as well as robustness against departures from narrow model classes. This has made them very popular in the last decades, and the attractiveness of nonparametric methods transfers to many data science applications where specific parametric models are not justifiable. However, a shortcoming of those inference procedures that rely on the nonparametric relative effect (Mann-Whitney functional) as their base functional is their inability to capture differences between distributions that cannot be described by a stochastic tendency. To this end, we have introduced a functional describing distributional overlap and derived a consistent estimator along with its asymptotic distribution, even jointly with that of the relative effect estimator. Combining these two functionals allows for much more versatile inference, which we will demonstrate in this presentation. Also, we will try to address the issue of interpretability of the resulting effect measures, as straightforward interpretability is key to their usability in practice.

H0205: **Functional profile completion: An R package**
Presenter: **Matus Maciak**, Charles University, Czech Republic

Risk assessment is typically based on various parametric approaches using rather strict theoretical assumptions, but real data often do not comply with these restrictions. Alternatively, non-restrictive and easily applicable nonparametric functional-based techniques relying on completion of missing profiles can be used. We present the first R package that fully implements such nonparametric risk assessment techniques. Three competitive algorithms are discussed in particular. A brief theoretical background and important statistical properties are addressed together with some practical illustrations and real datasets.

H0179: **Functional shape outliers as contamination in complexity mixtures**
Presenter: **Enea Bongiorno**, Universita del Piemonte Orientale, Italy
Co-authors: Aldo Goia, Kwo Lik Lax Chan

Shape outliers are treated as contamination elements and part of a high-complexity component within an appropriate mixture model of functional data. The aim is threefold. First, we define the notion of complexity based on the concept of small ball probability. Second, we theoretically introduce the idea of a complexity mixture and analyze its implications on small ball probabilities. Third, we propose an algorithm to decompose a complexity mixture into its constituent components, thereby implicitly identifying potential contamination in a functional dataset. The effectiveness of the proposed methodology is demonstrated through an application.

H0197: **Generalization of the Mahalanobis distance for star-shaped sets: An application to fuzzy clustering**
Presenter: **Ana Belen Ramos-Guajardo**, Fundacion Universidad de Oviedo, Spain
Co-authors: Maria Brigida Ferraro, Gil Gonzalez-Rodriguez

Several clustering techniques for imprecise information have emerged over the past few decades. Some of these methods incorporate fuzziness into the clustering process, such as the widely recognized fuzzy k-means algorithm. This algorithm has also been previously developed to handle the clustering of star-shaped sets. However, the fuzzy k-means method has a limitation: it does not account for the correlation structure between variables, which becomes problematic when the clusters are not spherical in shape. To overcome this drawback, the Mahalanobis distance, wich considers covariance matrices between variables, has been introduced. Thus, a novel fuzzy clustering algorithm for star-shaped sets based on the Mahalanobis distance is proposed. The performance of both the fuzzy k-means method and the proposed approach is evaluated through a real-world application.

HO003 Room Amphitheater 2 DIMENSION REDUCTION Chair: Andreas Artemiou

H0158: **Sufficient dimension reduction for the conditional quantiles of functional data**
Presenter: **Eliana Christou**, University of North Carolina at Charlotte, United States
Co-authors: Eftychia Solea, Shanshan Wang, Jun Song

Functional data analysis holds transformative potential across fields but often relies on mean regression, with limited focus on quantile regression. Furthermore, the infinite-dimensional nature of the functional predictors necessitates the use of dimension reduction techniques. Therefore, in this work, we address this gap by developing dimension reduction techniques for the conditional quantiles of functional data. The idea is to replace the functional predictors with a few finite predictors without losing important information on the conditional quantile while maintaining a flexible nonparametric model. We derive the convergence rates of the proposed estimators and demonstrate their finite sample performance using simulations and a real dataset from fMRI studies.

H0184: **Robust inverse regression for multivariate elliptical functional data**
Presenter: **Eftychia Solea**, Queen Mary University of London, United Kingdom

Functional data have received significant attention as they frequently appear in modern applications, such as functional magnetic resonance imaging (fMRI). The infinite-dimensional nature of functional data makes it necessary to use dimension reduction techniques. Most existing techniques, however, rely on the covariance operator, which can be affected by heavy-tailed data and unusual observations. Therefore, we consider a robust functional sliced inverse regression (R-FSIR) for multivariate elliptical functional data. We define the elliptical distribution for a vector of random functions. We introduce a new statistical linear operator, called the conditional spatial sign Kendall's tau covariance operator, which can be seen as an extension of the multivariate Kendall's tau to both the conditional and functional settings, and is capable of handling heavy-tailed functional data and outliers. We show that the conditional spatial sign Kendall's tau covariance operator has the same eigenfunctions as the conditional covariance operator. Hence, we can formulate the generalized eigenvalue problem based on this new operator to achieve estimation robustness. We derive the convergence rates of the proposed estimators for both completely and partially observed data. Finally, we demonstrate the finite sample performance of our estimator using simulation examples and a real dataset based on fMRI.

H0195: **Robust sufficient dimension reduction for multivariate time series analysis**
Presenter: **Amal Alqarni**, Cardiff University, United Kingdom

Sufficient dimension reduction (SDR) is a statistical framework that is widely used to reduce high-dimensional data while maintaining significant information. It has applications in a wide range of domains. The previous methodology of SDR in the context of multivariate time series analysis (like TSIR) is extended, addressing complex challenges such as high dimensionality, temporal relationships, and noise arising from heavy-tailed outliers. We propose a novel methodology for robust dimension reduction in multivariate time series analysis: Time-series Sliced Inverse Median Difference (TSIMeD). TSIMeD achieves robust dimension reduction by identifying key temporal directions and lags, even in the presence of heavy-tailed outliers. To enhance dimension selection, we propose a Bayesian Information Criterion (BIC)-type method, improving model interpretability and efficiency. Extensive simulations demonstrate that TSIMeD outperforms established methods like Time Series Sliced Inverse Regression (TSIR) and Sliced Inverse Mean Difference (TSIMD), requiring fewer directions while maintaining superior accuracy across a variety of lag settings and noise conditions, highlighting their transformative potential for multivariate time series analysis.

H0156: **Unveiling the latent space: Dimension reduction for concepts discovery in GNN-based combinatorial optimization**
Presenter: **Havana Rika**, The Academic College of Tel Aviv - Yaffo, Israel
Co-authors: Dan Vilenchik, Joowon Lee, Joowon Lee

In recent years, the intersection of machine learning and optimization of NP-hard problems has been marked by significant advances in the application of Graph Neural Networks (GNNs). Despite these successes, fundamental questions remain about how and why these models make their decisions. In classical combinatorial optimization, experts have long relied on interpretable heuristics (such as notions of vertex support or degree) to guide problem-solving. By contrast, neural networks typically operate as black boxes, obscuring the intuitive concepts that may emerge within their high-dimensional latent spaces. Explainable AI (XAI) methods provide a pathway to peer inside the network and identify the ideas that drive performance. Among the diverse XAI techniques, the dimension reduction approach known as Principal Component Analysis (PCA) has proven invaluable for revealing the low-dimensional structures that underlie a GNNs learned combinatorial concepts. The aim is to explore the insights and learned concepts of GNNs applied to solving NP-hard problems such as Boolean satisfiability (SAT), Graph Coloring, and the Max-Clique problem while comparing them to known combinatorial heuristics.

HO016 **Room Lecture room 4** STATISTICS FOR CORPORATE SUSTAINABILITY AND FINANCIAL RESILIENCE **Chair: Alessandra Amendola**

H0169: **Multiobjective optimization of ESG bond portfolios: A copula-based dynamic Nelson-Siegel approach**
Presenter: **Andreas Stephan**, Linnaeus University, Sweden
Co-authors: Maziar Sahamkhadam

A copula-based pricing framework is presented for forecasting bond returns and optimizing multiobjective bond portfolios (MOBPs). Utilizing a copula-based dynamic factor model, we generate step-ahead forecasts for zero-coupon bond yields, which are applied to price both callable and non-callable fixed-coupon bonds. These simulated bond prices serve as inputs for convex multiobjective portfolio optimization, incorporating key criteria such as average returns, Conditional Value-at-Risk (CVaR), distance-to-default, transaction costs, and option-adjusted duration and convexity. Applying our methodology to a dataset of 879 environmental, social, and governance (ESG) bonds denominated in Euros from January 2016 to July 2024, we demonstrate that the proposed MOBP approach consistently outperforms an equally weighted (EQW) benchmark in terms of higher returns and Sharpe ratios while effectively mitigating tail risk. Notably, our framework improves portfolio resilience during market turbulence, such as the COVID-19 pandemic and the Russo-Ukrainian war, underscoring its applicability in risk-sensitive sustainable investing.

H0175: **Measuring inequality in the adoption of ESG scores by small and medium enterprises**
Presenter: **Alessandra Amendola**, University of Salerno, Italy
Co-authors: Paolo Giudici, Adelaide Emma Bernardelli

Inequality has been a central focus in discussions on sustainable development, and it is typically measured using metrics such as the Gini index and the Lorenz Curve. While these measures were initially applied to income inequality, they have progressively been used to assess disparities in access to essential resources, environmental impact, and infrastructure. However, little attention has been given to inequality in Environmental, Social, and Governance (ESG) performance, which is vital for both firm strategies and market dynamics. Existing ESG studies focus on its financial benefits and performance, but the uneven distribution of ESG adoption, particularly among small and medium-sized enterprises (SMEs), remains underexplored. This issue is particularly relevant following Italy's adoption of the Corporate Sustainability Reporting Directive (CSRD), which mandates ESG reporting for large enterprises and listed SMEs. The aim is to introduce a novel methodology to assess ESG adoption among Italian SMEs, measuring inequality in ESG scores as a proxy for diffusion. The dataset, which contains ESG scores and a set of financial variables, is analyzed using the Gini index to predict ESG disparities and assess model performance. Additionally, the study extends the analysis with multivariate approaches, incorporating linear regressions and neural networks to explore the determinants of ESG inequality.

H0190: **Structural patterns and country effects in global ESG performance**
Presenter: **Marialuisa Restaino**, University of Salerno, Italy
Co-authors: Michele La Rocca, Marcella Niglio, Steven Mphaya

The relationship between Environmental, Social, and Governance (ESG) performance across firms from multiple countries is investigated. Particular attention is focused on identifying and quantifying the significant influence of country-specific factors. In more detail, using firm-level ESG data from a diverse set of countries, the study focuses on identifying patterns and disparities in ESG scores across firms from different countries, highlighting whether and how significant country effects that influence the distribution of the ESGs exist. Employing some statistical analysis on a dataset of companies, we aim to identify key patterns, distributions, structural differences in ESG metrics, and distinct groups of firms exhibiting similar ESG profiles. Moreover, by comparing the analysis across different countries and incorporating country-level variables, this analysis seeks to disentangle the impact of national contexts on the observed relationships. The findings will provide valuable insights for investors, policymakers, and corporate managers seeking to understand the role, dynamics and importance of ESG integration. They will also highlight the importance of considering country-specific characteristics in ESG analysis and decision-making.

H0194: **Predicting bankruptcy of micro-enterprises by industry: Integrating financial and web-based indicators**
Presenter: **Caterina Liberati**, University of Milano-Bicocca, Italy
Co-authors: Carlo Bottai, Lisa Crosato

The aim is to study bankruptcy prediction for micro-sized enterprises, often underrepresented in credit risk modeling due to their limited financial data quality. Building on previous research advocating for sector-specific approaches, we develop separate prediction models for selected industries using a dataset of 84,019 Italian micro-enterprises, of which only 1,308 (1.15%) defaulted. The low default rate makes the classification problem particularly complex, especially when analyzed by sector. To overcome the limitations of models based solely on balance sheets, we integrate an innovative non-financial information source: features extracted from the HTML structure of company websites. These web-based indicators are combined with traditional financial variables to enhance model performance. A cross-validation scheme ensures the robustness and generalizability of results. Findings show that website data add significant predictive power, particularly in industries where digital presence is actively maintained. The relevance of these features varies across sectors, underlining the presence of sector-specific heterogeneity not only in financial patterns but also in web behavior. We demonstrate that website information represents a valuable and innovative signal for early-warning systems, especially in data-poor environments. This approach offers new perspectives for more accurate and industry-aware credit risk models for micro-enterprises.

Wednesday 09.07.2025 16:10 - 17:25 Parallel Session I – HiTECCoDES2025

HC028 Room Amphitheater 1 ECONOMETRIC THEORY Chair: Masayuki Hirukawa

H0221: **Inference on panel data models with a generalized factor structure**
Presenter: **Juan Manuel Rodriguez-Poo**, Universidad de Cantabria, Spain
Co-authors: Alexandra Soberon, Stefan Sperlich

The focus is on the identification, inference, and validation of linear panel data models when a nonparametric function accounts for both factors and factor loadings. This general specification encompasses rather popular models such as the two-way fixed effects and the interactive fixed effects ones. By applying a conditional mean independence assumption between unobserved heterogeneity and the covariates, we provide consistent estimators of the parameters of interest at the optimal rate of convergence, for fixed and large T. We also provide a specification test for the modeling assumption based on the methodology of conditional moment tests and nonparametric estimation techniques. Using degenerate and nondegenerate theories of U-statistics, we show its convergence and asymptotic distribution under the null, and that it diverges under the alternative at a rate arbitrarily close to $\sqrt{NT}$. Finite sample inference is based on bootstrap. Simulations reveal an excellent performance of our methods, and an empirical application is conducted.

H0219: **Estimation of functional coefficient panel data models with endogenous selectivity and fixed effects**
Presenter: **Alexandra Soberon**, Universidad de Cantabria, Spain
Co-authors: Daniel Henderson, Juan Manuel Rodriguez-Poo, Taining Wang

A novel estimation approach is developed for functional coefficient panel data models with sample selection and fixed effects. We propose a two-step pairwise approach that avoids strict identification restrictions and addresses individual heterogeneity and selection bias. The first stage estimates the selection equation parameters, while the second stage estimates the regression of interest using a generalized local weighting scheme that removes the sample selection bias asymptotically using the estimates of the previous stage. We establish the asymptotic properties of the proposed estimators under rather weak assumptions and demonstrate the method's superior computational efficiency with respect to existing approaches and finite-sample performance through Monte Carlo simulations.

H0170: **NA-DiD: Extending Difference-in-Differences with capabilities**
Presenter: **Stanislaw Halkiewicz**, AGH University of Cracow, Poland

A novel reinterpretation is proposed for the classical Difference-in-Differences (DiD) estimator as an integral operator, where treatment effects are aggregated over time using an underlying measure. In the standard framework, the DiD coefficient can be expressed as a Lebesgue integral, which implicitly assumes linearity and additivity in the contribution of treatment periods. Building on this insight, we introduce the Non-Additive Difference-in-Differences (NA-DiD) framework, which generalises the aggregation scheme by replacing the additive measure with a capacity and employing the Choquet integral. This generalisation allows for modelling non-linear aggregation patterns, including synergies, diminishing returns, and threshold effects, which are often observed in real-world interventions but cannot be accommodated within the classical model. The NA-DiD estimator preserves the structure of DiD while relaxing its restrictive assumptions, and coincides with the classical form under additivity. We demonstrate the usefulness of the approach using a simulated intervention with time-varying treatment intensity, where NA-DiD yields more conservative and temporally nuanced estimates than its classical counterpart. The proposed framework provides a flexible and theoretically coherent extension of DiD, particularly well-suited for evaluating interventions with dynamic or non-monotonic effects that challenge the assumptions of standard causal inference methods.

HC030 Room Amphitheater 2 APPLIED STATISTICS AND ECONOMETRICS Chair: Maria Brigida Ferraro

H0217: **A partially pooled Bayesian hierarchical model for aggregated relational survey data**
Presenter: **Rowland Seymour**, University of Birmingham, United Kingdom

Aggregated relational data can be collected in household surveys to estimate the number of people who have been affected by crimes. This method asks survey respondents questions of the form 'How many people do you know with feature X?'. Surveys of this kind can require large sample sizes when the social network structure of the population is heterogeneous. To allow us to reduce the sample size and simultaneously provide estimates for different subgroups in the population, we develop a partially pooled Bayesian hierarchical model. Through a linear predictor, we introduce correlation between the subgroup model parameters and assume that the parameters for the subgroups come from a national-level distribution, which allows us to share information between the subgroups. Inference for the model can be quickly implemented in Stan. We demonstrate this model on a new data set on child sexual abuse in the Philippines, and show how the results led to new laws in the USA and the Philippines.

H0226: **Financial literacy and advice: Substitutes or complements**
Presenter: **Demetris Koursaros**, Cyprus University of Technology, Cyprus

The aim is to investigate the relationship between financial literacy and financial advice. We first introduce a simple two-period convincing model between a financial advisor and a household. In our model, households have heterogeneous beliefs with respect to the distribution of true returns. More dispersed beliefs correspond to less sophisticated households. We show that financial advisors face a lower cost to shift beliefs and convince a financially unsophisticated household to invest in a larger amount, which is, however, a sub-optimal decision for the household, and earn a higher reward. Eventually, more financially sophisticated households are more likely to ask for financial advice. We extend our model in many directions and provide simulation exercises. A novelty of our model is that we allow the sophistication level of the advisor to vary and provide empirical predictions for the matching with households with different sophistication levels. We test these within a large sample of Canadian financial advisors and their clients.

H0227: **Long-term forecasting of stock returns: Avoid overly complex machine learning and prioritize benchmarking**
Presenter: **Parastoo Mousavi**, Bayes Business School, City St Georgeś, University of London, United Kingdom
Co-authors: Jens Perch Nielsen, Tatiana Franus

Machine learning is increasingly the default choice for data analysis, often regarded as the only solution. The value of incorporating simple, intuitive models is argued when forecasting long-term stock returns. We show that by focusing on manual optimization, especially in data-constrained environments, simpler models can outperform automated machine learning methods. Our findings highlight the critical role of human oversight in financial forecasting problems and challenge the assumption that automated approaches always deliver superior results.

Authors Index