ISBIS 2023 Satellite Conference

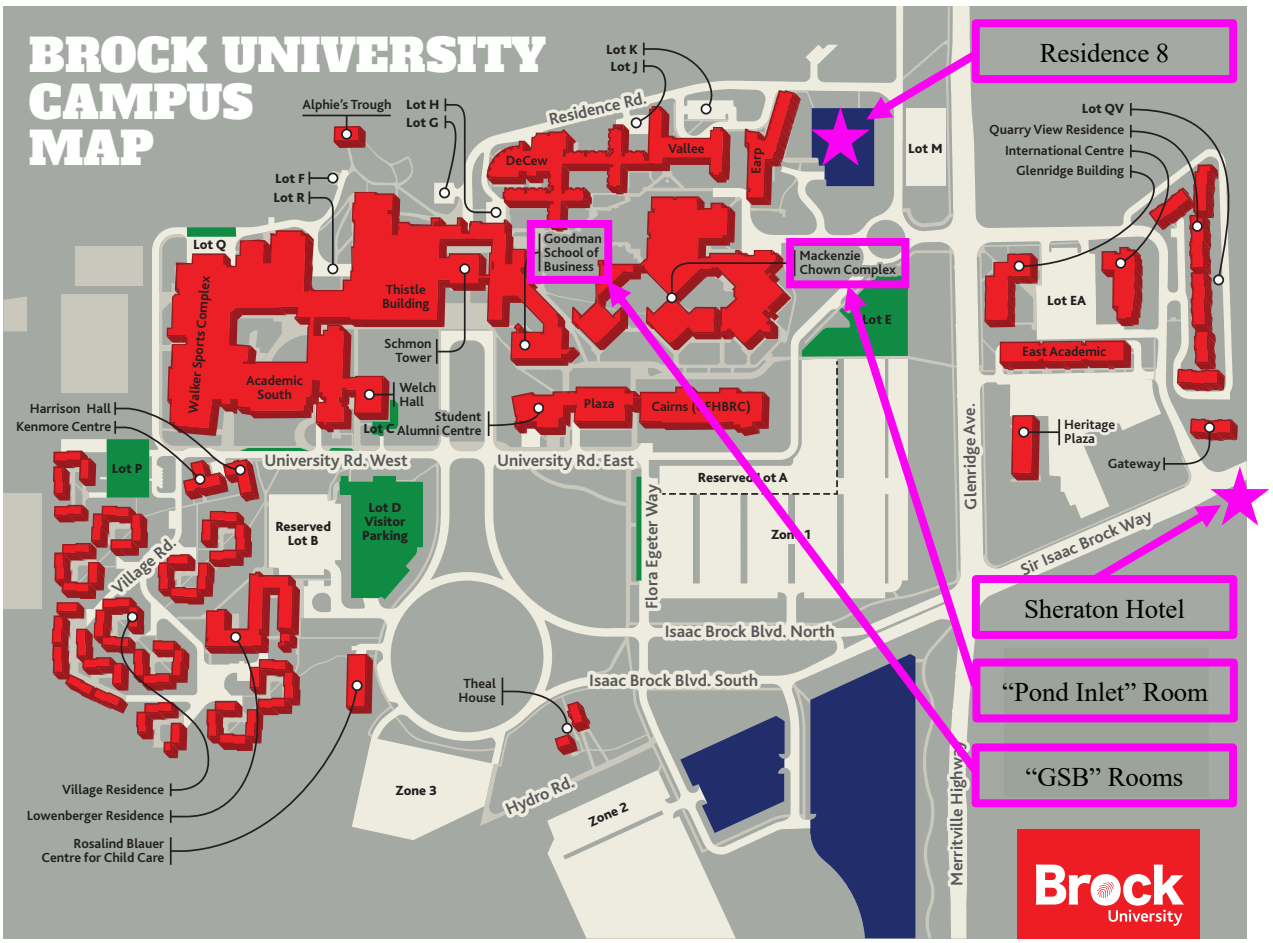
ISBIS INTERNATIONAL SOCIETY FOR BUSINESS AND INDUSTRIAL STATISTICS

SCHEDULE: DAY 1

	THURSDAY JULY 13, 2023	
8:00 – 9:00	**Breakfast & Registration (Rm: Pond Inlet)**	
9:00 – 9:15	**Welcome & Opening Remarks (Rm: Pond Inlet)**	
9:15 –10:25	**Keynote Address (Rm: Pond Inlet)** **"A Data Scientist Reads the News" – Bonnie Ray, ChartBeat**	
10:25 – 10:50	**Coffee Break (Rm: GSB 306)**	
10:50 – 12:20	**Invited Session 1 (Rm: GSB 307)** **Deep Learning**	**Contributed Session 1 (Rm: GSB 308)** **Advances in Statistical Modeling I**
12:20 – 13:30	**Lunch (Rm: Pond Inlet)**	
13:30 – 15:00	**Invited Session 2 (Rm: GSB 307)** **High Dimensional Data Analysis**	**Contributed Session 2 (Rm: GSB 308)** **Advances in Statistical Modeling II**
15:00 – 15:20	**Coffee Break (Rm: GSB 306)**	
15:20 – 16:50	**Invited Session 3 (Rm: GSB 307)** **Sports Analytics I: Soccer**	**Contributed Session 3 (Rm: GSB 308)** **Advances in Clinical Studies**
16:50 – 17:00	**Break (Rm: GSB 306)**	
17:00 – 18:00	**Invited Session 4 (Rm: GSB 307)** **Sports Analytics II: Basketball & Golf**	**Invited Session 5 (Rm: GSB 308)** **Online Experiments and A/B Tests**
18:00 – 18:30	**Break**	
18:30	**Banquet (Rm: Pond Inlet)**	

SCHEDULE: DAY 2

	FRIDAY JULY 14, 2023	
8:00 – 8:30	**Breakfast & Registration (Rm: Pond Inlet)**	
8:30 – 10:00	**Invited Session 6 (Rm: GSB 307)** **Advances in Univariate Time Series**	**Invited Session 7 (Rm: GSB 308)** **ENBIS Session**
10:00 – 10:30	**Coffee Break (Rm: GSB 306)**	
10:30 – 12:00	**Invited Session 8 (Rm: GSB 307)** **Advances in Time Series**	**Invited Session 9 (Rm: GSB 308)** **ISTAT & SDS-SIS Session**
12:00 – 13:00	**Closing Remarks & Boxed Lunch To Go (Rm: Pond Inlet)**	



ABSTRACTS

KEYNOTE ADDRESS (9:15 – 10:25, JULY 13)
Room: Pond Inlet

A data scientist reads the news: What we can learn about the state of the world through analysis of the news and reader habits

Bonnie Ray (ChartBeat)

Understanding reader engagement with digital content is critical not only for driving editorial decisions and revenue strategies for news publishers, but also for gaining a broader perspective on the topics that are resonating with readers across a region and the world. In this talk, I'll discuss some of the statistical and computational challenges faced at Chartbeat, a content analytics company that works with publishers in over 70 countries measuring over 60B pageviews per month. In particular, I'll discuss the development of our pipeline for scraping, extracting, and linking key topics from articles published by our customers in many different languages. I'll then show examples of how we use this data, together with measures of reader engagement, to gain insight into which stories are driving the most reader attention, what behaviors lead to registrations and subscriptions, and an approach to article recommendations that tries to balance topic similarity with popularity, timeliness, and diversity of thought. I'll end with some thoughts around how state-of-the-art AI capabilities stand to impact the news and media industry.

INVITED SESSION 1 (10:50 – 12:20, JULY 13)
DEEP LEARNING
Organizer/ Chair: Farouk Nathoo (University of Victoria)
Room: GSB 307

Online Kernel-Based Mode Learning for Big Data

Tao Wang (University of Victoria)

Data online modeling is an important research direction in the fields of economics, statistics, machine learning, and data science. In practical applications, big data with an exceptionally large sample size are frequently containing outliers or following heavy-tailed distributions. An online learning estimation with anti-outlier capabilities is therefore urgently required to achieve robust and efficient estimators. Instead of the conventional least squares or Huber loss function, we in this paper introduce an online learning strategy built on the mode kernel loss function to account for outliers and heavy-tailed distributions in big data. In order to achieve robust estimators and reduce computational complexity with a minimum loss of statistical efficiency, we incorporate mode regression into an online learning structure with subsets of data, which can continuously update historical data using important features acquired from new data subsets. By merging asymptotic distribution functions generated from all local mode regression estimators, the newly suggested estimator can be efficiently updated as the minimizer of a weighted least squares type loss function. With the aid of a normal kernel function, a modified modal expectation-maximization algorithm is developed to numerically solve the model. Under a general likelihood estimation framework, the resultant estimator is demonstrated to be asymptotically equivalent to the estimator calculated using the entire data when the covariates of each subset are homogeneous. Monte Carlo simulations and an empirical study pertaining to American airlines are presented to illustrate the finite sample performance of the proposed estimator.

Integrating Neural Networks into Functional Data Analysis

Cédric Beaulac (Université du Québec à Montréal)

In this presentation, I will introduce the recent findings in the research program I am developing with my students about the integration of neural networks and machine learning models in the analysis of functional data. As a first step, we propose a solution to the function-on-scalar regression problem: a novel neural network layer architecture. The proposed functional output layer allows us to use neural networks to output functional responses. To do so, the second-to-last layer is designed to output basis coefficients that are combined in the last layer with associated basis functions in order to output a functional response. We also designed a roughness penalty that can be integrated in the optimization

process of the proposed functional neural network as a regularizer. Second, we propose a concept for functional weights which can project functional data to a scalar representation and can accommodate irregularly spaced data. Combining the proposed functional weights with the functional output layer led to the design of a functional autoencoder. This autoencoder is built to return a finite representation of the functional data and a smooth reconstruction of that data simultaneously. This model can be described as a non-linear functional principal component analysis. Next, we showcase possible applications of these models using real data. We will demonstrate the benefit they provide and discuss when these techniques should be considered. To conclude, we will introduce implementations of these models that we made publicly available.

Feature Extraction using a Neural Network Classifier
Farouk Nathoo (University of Victoria)

A major issue in modern regression problems is the association of images to high-dimensional covariates. A biomedical example is the association of genes to neuroimaging phenotypes, though this problem is relevant to other areas as well where both the response and covariates are high-dimensional. In this article, we tackle the latter problem with an eye toward developing solutions that are relevant for disease prediction. Supported by a vast literature on the predictive power of neural networks, our proposed solution uses neural networks to extract from neuroimaging data features that are relevant for predicting Alzheimer's Disease (AD) for subsequent relation to genetics. Our regression analysis pipeline is comprised of image processing, neuroimaging feature extraction and genetic association steps. We propose a neural network classifier for extracting neuroimaging features that are related with disease and a multivariate Bayesian group sparse regression model for association. We compare the predictive power of these features to expert selected features and take a closer look at the variables identified. while the problem is formulated with eye towards a biomedical problem, potential applications to problems in business or finance are of great interest.

INVITED SESSION 2 (13:30 – 15:00, JULY 13)
HIGH DIMENSIONAL DATA ANALYSIS
Organizer/ Chair: Ejaz Ahmed (Brock University)
Room: GSB 307

Inference for High Dimensional Models: Linear Regression and Censored Quantile Regression
Yi Li (University of Michigan)

Drawing inferences in high-dimensional models is difficult due to the limitations of regular asymptotic theories. A new framework, called Selection-assisted Partial Regression and Smoothing (SPARES), is presented to address this challenge. SPARES reduces the high-dimensional problem to low-dimensional least squares estimations by using data splitting, variable selection, and partial regression. The method is extended to handle censored quantile regression models, which are widely used to model the heterogeneous effects of biomarkers on censored outcomes. The SPARES estimator is shown to be consistent and asymptotically normal, with its variance derived through a non-parametric delta method. The effectiveness of the method is demonstrated through simulations and comparisons with de-biased LASSO estimators. The SPARES method is applied to two genomic datasets, resulting in biologically meaningful results.

Sparse estimation in Markov regime-switching models
Abbas Khalili (McGill University)

Markov regime-switching vector auto-regressives are frequently used for modelling heterogeneous and complex relationships between variables in multivariate time series analysis. Applications include analyzing macroeconomic time series such as manufacturing activities, consumer price indices, and housing and asset prices. The most common method of estimation in these models is maximum likelihood estimation (MLE). However, even for moderate data dimension and number of regimes, the MLE becomes unstable. In this talk, we present regularization-based estimators when the number of regimes in the model is correctly or over-specified. We also discuss theoretical and finite-sample performances of the methods, including forecasting, and conclude with a real data analysis.

Improving the prediction accuracy in high-dimensional data analysis
Ejaz Ahmed (Brock University)

Abstract: In high-dimensional scenario where the number of predictors is greater than observations, many penalized techniques are available for simultaneous variables selection and parameters estimation under the model sparsity assumption. However, in a host of investigations a model may have sparse signals together with a number of weak signals. In such cases variable selection procedures may not be able to differentiate predictors with weak signals and sparse signals. Thus, the prediction based on a selected submodel may not be desirable. For this reason, we propose a high-dimensional shrinkage strategy to improve the prediction performance of a submodel generated from most existing variable selection methods. Such a high-dimensional shrinkage estimator (HDSE) is constructed by shrinking a weighted ridge estimator in the direction of a candidate submodel. We demonstrate that the proposed HDSE performs uniformly better than the weighted ridge estimator. Interestingly, it improves the prediction performance of given submodel drastically. The relative performance of the proposed HDSE strategy is appraised by both simulation studies and the real data analysis. Some open research problems will be discussed, as well.

INVITED SESSION 3 (15:20 – 16:50, JULY 13)
SPORTS ANALYTICS I: SOCCER
Organizer/ Chair: Jean-François Plante (HEC Montréal)
Room: GSB 307

Comparison of Individual Playing Styles in Soccer
Tianyu Guan (Brock University)

This research attempts to identify soccer players who have a similar style to a player of interest. Playing style is not adequately quantified with traditional statistics, and therefore style statistics are created using tracking data. Tracking data allow us to monitor players throughout a match, and therefore includes both “on-the-ball” and “off-the-ball” observations. Having developed style statistics, tractable discrepancy measures are introduced that are based on Kullback-Leibler divergence in the context of multivariate normal distributions. An example is provided where a pool of players from the Chinese Super League are identified as having a playing style that is similar to Marouane Fellaini.

Applications of Poisson models for match outcomes in Soccer
Alexandre Leblanc (University of Manitoba)

In this presentation, we will discuss the use of Poisson models for match outcomes in professional soccer. Beyond their simplicity, these models can be used to answer a variety of questions that are interesting to coaches and fans of the game alike. Some of our investigations around such questions will be presented. We will also briefly discuss shortcomings of basic Poisson models and some of the useful fixes and generalizations.

Sports Analytics: Data & Modelling Challenges when Working with Match Data
Cristina Rizzuto (DMX Analytics)

Sports analytics data was traditionally limited to match event data (passes, interceptions, shots, etc.). With the arrival of tracking data, we now have access to all player and ball positions (X,Y) 10 times per second during matches; several new analysis opportunities are now available. However, there are also multiple difficulties when working with match data and we will present some of them, such as the differences in variables when using different data providers and the impact on modelling. Using soccer data as an example, we will also demonstrate how tracking data can be transformed in useful features for a passing model and an expected goals model. Furthermore, we will discuss how our models can provide tactical insights that are helpful to a coaching staff. We will highlight modelling approaches that have been successful and present some of the improvement opportunities that remain.

INVITED SESSION 4 (17:00 – 18:00, JULY 13)
SPORTS ANALYTICS II: BASKETBALL & GOLF
Organizer: Nathaniel Stevens (University of Waterloo)
Chair: Alexandre Leblanc (University of Manitoba)
Room: GSB 307

NBA Lineup Attribution via Spectral Analysis
Steve Devlin (University of San Francisco)

Winning in the NBA requires good players who play well together. But how can the synergistic effects (or lack thereof) of player groups be quantified? Is a big-three really more than the sum of its parts? We address the problem of player group attribution using spectral analysis, a novel approach from algebraic signal processing. Just as a time series can be decomposed into contributions in the frequency domain, we leverage the algebraic structure of team lineups to decompose the team success signal into components that naturally correspond to player group contributions. We present a detailed a spectral analysis of NBA data to show how it can be a practical tool for use in lineup evaluation.

What can golf teach us about humans?
Jean-François Plante (HEC Montréal)

Sports Analytics leverages data to learn about the game, the teams, the fans, as well as the business of sports. But sports data may also teach us about the humans that are involved. The PGA ShotLink data contains a detailed record of professional golf covering multiple years. With different colleagues, we used it to provide empirical evidence to answer various questions. Do people react differently to glory vs. money? How do they modulate their risk when under pressure? Are foreigners at a disadvantage when they enter the PGA? Who among the newcomers are more likely to become successful? I will present an overview of the results of papers that answer these questions and more.

INVITED SESSION 5 (17:00 – 18:00, JULY 13)
ONLINE EXPERIMENTS & A/B TESTS
Organizer/Chair: Nathaniel Stevens (University of Waterloo)
Room: GSB 308

An Overview of Statistical Challenges in Online Controlled Experiments
Nathaniel Stevens (University of Waterloo)

The rise of internet-based services and products in the late 1990's brought about an unprecedented opportunity for online businesses to engage in large scale data-driven decision making. Over the past two decades, organizations such as Airbnb, Alibaba, Amazon, Baidu, Booking, Alphabet's Google, LinkedIn, Lyft, Meta's Facebook, Microsoft, Netflix, Twitter, Uber, and Yandex have invested tremendous resources in online controlled experiments (OCEs) to assess the impact of innovation on their customers and businesses. Running OCEs at scale has presented a host of challenges requiring solutions from many domains. In this talk we discuss the practice and culture of online experimentation, and we review practical challenges that require new statistical methodologies to address them. The goal is to raise academic statisticians' awareness of these new research opportunities so as to increase collaboration between academia and the online industry.

Hidden Integration Costs in Online Controlled Experimentation Platforms
Nick Ross (University of Chicago)

Data Science Practitioners have an array of options when implementing an AB-testing or Online Controlled Experiment system. At the first level is the "build" vs. "buy" decision with all its traditional trade-offs. If "buy" is chosen there are a multitude of solution providers (Google Analytics, Split, Amplitude, etc.) who compete on both costs and features. A relatively overlooked aspect of this decision is how the success or failure of an AB-testing initiative is intertwined with an organization's current data infrastructure. Organizations frequently underestimate the total costs to integrate with a third-party system and, in doing so, can end up choosing a suboptimal option. These costs can include engineering time, a loss of product features, time spent importing and exporting data as well as time lost to pursuing

down data rabbit. In this talk we will present a framework for analyzing integrations which will allow us to understand and predict issues before they arise and preemptively avoid them.

INVITED SESSION 6 (8:30 – 10:00, JULY 14)
NEW DEVELOPMENTS IN TIME SERIES ANALYSIS AND FORECASTING
Organizer: Paulo Rodrigues (Federal University of Bahia)
Chair: Nalini Ravishanker (University of Connecticut)
Room: GSB 307

This session was unfortunately cancelled

INVITED SESSION 7 (8:30 – 10:00, JULY 14)
ADVANCES IN STATISTICS FOR BUSINESS AND INDUSTRY (ENBIS)
Organizer/ Chair: Antonio Lepore (University of Naples)
Room: GSB 308

Escalator Health Analytics and Sensor Data Monitoring using LSTM Based control charts
Inez Zwetsloot (University of Amsterdam)

MTR, the major Hong Kong public transport provider, has been operating for 40 years with more than 1000 escalators in the railway network. These escalators are installed in various railway stations with different ages, vertical rises and workload. An escalator's refurbishment is usually linked with its design life as recommended by the manufacturer. However, the actual useful life of an escalator should be determined by its operating condition which is affected by runtime, workload, maintenance quality, vibration etc., rather than age only. Escalators in the same station are usually of the same age. Under the "time-based" strategy, escalators need to be refurbished more or less at the same time. This will inevitably cause inconvenience to the passengers and hence affect the level of service. If the refurbishment work is postponed without fully understanding the health condition of the escalators, the escalators may not operate well. The objective of this project is to develop a comprehensive health condition model for escalators to support the refurbishment decision. The analytic model consists of four parts: 1) online data gathering and processing; 2) condition monitoring; 3) a health index model; and 4) a remaining useful life model. The results can be used for 1) predicting the remaining useful life of the escalators, in order to support asset replacement planning and 2) monitoring the real-time condition of escalators; including signaling when vibration exceeds the threshold and signal diagnosis, giving an indication of possible root cause (components) of the signal. To develop the model the following data sources are utilized and combined: real-time vibration signals from eight sensors, continuous energy usage, as well as fault and maintenance history. In this talk, we will provide a short overview of this project and give details regarding the control charts developed for monitoring the real-time condition of escalators based on long short-term memory (LSTM) artificial neural networks.

A Test of Independence for Locally Stationary Processes with Application to Bridge Monitoring
Carina Beering (Helmut-Schmidt Universität)

We propose a testing procedure for independence of locally stationary processes using a weighted distance composed of characteristic functions (CF) and its empirical version as a base. This distance makes use of the unique behavior of joint characteristic functions of independent processes. The distance covariance defined by Székely et al. (2007) and its use by Jentsch et al. (2020) inspired the essential idea of this concept. To be finally able to compile a testing procedure, we provide the needed results with the notion of the beneficial effects of a bootstrap analogue. Therefore, we establish the bootstrap versions of the previously presented findings. Prior to that, we transfer the concept of the empirical weighted CF-distance to the bootstrap world. After some simulation studies to illustrate the functionality of our test regarding underlying independence as well as dependence of different forms, we apply our procedure to real bridge sensor data in the end. This data originates from structural health monitoring to identify dependencies between different sensor outputs.

The COVID19 impact of Italian firms' employment
Matilde Bini (European University of Rome)

This study aims to investigate the impact of COVID19 on employment for different types of Italian firms. The analysis is carried out using a double step methodology, the propensity score matching and the Dif-in-Dif model. The "treatments" are represented, in this case, by the different typologies of firms as registered in the pre-COVID19 period, and the characteristics of firms we considered are the innovation capacity, the public subsidies in the COVID19 period, the occupational mix in terms of "flexible" and "non-flexible" employee. The results we obtain confirm that Covid affected on employment according to the different typologies firms. This is joint work with Alessandro Zeli and Leopoldo Nascia.

INVITED SESSION 8 (10:30 – 12:00, JULY 14)
MODELING TIME SERIES WITH INTERESTING DEPENDENCE STRUCTURES
Organizer: Nalini Ravishanker (University of Connecticut)
Chair: Paulo Rodrigues (Federal University of Bahia)
Room: GSB 307

Multivariate Non-Linear Time Series with Spatial Considerations
Kathy B. Ensor (Rice University)

Of important consideration are multivariate nonlinear dynamic time series with low to high levels of spatial association. We explore a state-space hierarchical modeling approach, considering both a frequentist and Bayesian perspective. Key questions answered are natural clustering of the time series, short-term deviations between the series, and short-term predictions based on the fitted models. The methodology is applied to fifty weekly time series spanning three years, representing wastewater signals for SARS CoV-2. Wastewater signals are compared to the corresponding observed cases. From this paradigm, a predictive model for emergent diseases is posited.

Robust singular spectrum analysis: Comparison between classic and robust approaches for model fit and forecasting
Paulo Rodrigues (Federal University of Bahia)

Singular Spectrum Analysis (SSA) is a powerful and widely used non-parametric method to analyse and forecast time series. Although SSA has proven to outperform traditional parametric methods for model fit and model forecasting, one of the steps of the SSA algorithm is the singular value decomposition (SVD) of the trajectory matrix, which is very sensitive to the presence of outliers because it uses the L2 norm optimization. Therefore, the presence of outlying observations has a significant impact on the SSA reconstruction and forecasts, a problem that can be solved by considering robust methods. In this talk, I will give a general overview of SSA, will present some algorithms for the robust SSA, and will illustrate the results and comparisons, in terms of model fit and model forecasting, via Monte Carlo simulations based on synthetic and real data, considering several contamination scenarios. Joint work with Jonatha Pimentel, Patrick Messala, Mohammad Kazemi, Vanda Lourenço, and Rahim Mahmoudvand.

Sparse Multiplicative Error Models for Multivariate Positive-valued Financial Time Series
Nalini Ravishanker (University of Connecticut)

Univariate multiplicative error models (MEM) have been used for modeling positive-valued time series which are ubiquitous in finance for modeling financial variables such as realized volatilities, transaction volumes, durations, etc. In vector MEM's (vMEM), the variables of interest are modeled as the Hadamard product of a vector of conditional expectations and an error vector with unit mean and a positive definite covariance matrix, with several choices of error distribution. One complication with these vector models is that they are severely affected by the curse of dimensionality which arises due to the fact that the number of parameters grows quadratically with the increase in number of component series and lags. We discuss regularized estimation for modeling multivariate positive-value time series in the log-vMEM framework by modifying the approach discussed in recent literature for vector autoregression (VAR) models using three types of flexible and interpretable hierarchical lag (HLag)

structures. We discuss penalizing the likelihood using convex group lasso as well as non-convex penalties such as group smoothly clipped absolute deviation (SCAD) and group minimax concave penalty (MCP). We formulate the penalized likelihood such that the optimization algorithm can be run in parallel across components. We demonstrate our approach on simulated data and illustrate using empirical data on five robust intraday realized volatility measures for Microsoft. Our approach can also be used to model any multivariate positive-valued time series exhibiting persistence, such as multivariate modeling of a realized volatility metric for several stock indices. This is joint work with Chiranjit Dutta (eBay) and Sumanta Basu (Cornell University).

INVITED SESSION 9 (10:30 – 12:00, JULY 14)
DATA SCIENCE AND OFFICIAL STATISTICS OF THE NIS AND SDS GROUPS OF THE ITALIAN STATISTICAL SOCIETY
Organizer/ Chair: Matilde Bini (European University of Rome)
Room: GSB 308

Evaluating the quality of the Register for Public Administrations through the Total Process Error

Orietta Luzi (National Institute of Statistics)

During the last decade, the Italian National Institute of Statistics (Istat) has been engaged in a modernization program involving the use of statistical registers integrated into a single logical environment, the Italian Integrated System of Statistical Registers (ISSR), for supporting the consistency of statistical processes and improving the quality of information for users. One object of the ISSR is the satellite REgister for Public Administrations (REPA) that contains information on structural and economic variables on a subset of the Italian Public Administrations (PA). REPA extends, for each unit, structural information coming from the base business register related to PA with some economic variables obtained by integrating administrative and survey data. REPA includes different sub-populations, such as local governments, ministries, constitutional bodies, social security funds, etc. Each sub-population has a peculiar structure and classification of its economic data: by properly aggregating a selection of items, the so-called variables Frame PA harmonized for all the Italian PA are obtained. The design and implementation of REPA has been completed for the sub-population of local governments. While in Istat there are consolidated systems for assessing the quality of statistical survey processes, an appropriate quality framework for multisource production processes such as statistical registers were needed. In particular, a Total Process Error (TPE) model has been recently proposed in literature as a flexible tool to assess the quality of multisource processes. TPE proposes two phases of analysis: (1) Assessment of Single Data Sources with respect to Original Source Purposes and (2) Combination/Re-Use/Integration of Data Sources with respect to Target Statistical Purposes. The second phase can be split into two subphases: (2a) Assessment of Single Data Sources with respect to Target Statistical Purposes and (2b) Assessment of the Combined Data Sources with respect to Target Statistical Purposes. In addition, TPE uses an operational tool to connect the steps of the production process to the phases of the quality evaluation framework. In this paper, we describe the application of TPE to the REPA local governments.

The new European business microdata source (MDE) for Trade statistics: asymmetries analysis via selective editing

Monica Pratesi (National Institute of Statistics)

In the new European Statistical System framework, aimed at sharing harmonized data, tools and methodologies with the purpose of improving data quality and reducing statistical burden on enterprises, a new statistical data source for measuring business globalization has been made available since 2022. The new data source consists in nationally collected export micro-data that are exchanged among Member States. They provide the receiving Member State with a detailed data source for the compilation of intra-EU imports, allowing the reduction of statistical burden on importers. Moreover, at the same time, the new data source gives the possibility to gain knowledge of the origin of asymmetries in trade data and to reconcile them, reducing the level of distortion in the statistical measure of globalization. In this framework, Istat is currently involved in the project “Development of methods and tools to analyze trade asymmetries in IT MDE data: detection of main discrepancies and related reconciliation actions”.

One of the main goals of the grant action is to develop a harmonized tool to detect and explain asymmetries in trade data, to be shared within the European Statistical System. This would allow the development of a common harmonized analysis approach. The method, implemented in an open-source code, is based on selective macro-editing. The strategy adopted allows the detection of the most relevant asymmetries in a top-down approach, analyzing the discrepancies between MDE data and nationally collected import data by country, by product and by trader. Once detected, the most relevant asymmetries are analyzed in detail, both exploiting available in-house statistical data sources, or by directly contacting the importers, prioritizing the asymmetries identified as persistent over the months. This strategy allows us to reduce the impact of trade asymmetries in disseminated statistics, as well as to assess potential quality issues in MDE data. It allows us to solve them and to integrate the new data source in the Intra-EU trade national production system. The results obtained so far demonstrate an improvement in efficiency, both in terms of detection of the most relevant discrepancies to be investigated and in terms of reduction of the micro-data to be checked in detail by manual revision.

An Integrated Unsupervised and Supervised Classification Strategy: an application to Credit Card Fraud Detection

Rosanna Verde (University of Campania Luigi Vanvitelli)

This work proposes two-step supervised classification strategy that combines unsupervised and supervised classification methods to improve the performance of classifiers. The proposal is applied to real data provided by a European financial partner for credit card fraud detection. It is performed in two stages: The first stage runs a clustering method, based on adaptive distance, to extract new knowledge from the data. This is achieved by identifying subclasses from a priori classes by optimising a discriminating criterion to consider changes in customer behaviour and the ability of fraudsters to invent new fraud patterns. Additionally, we introduce a feature selection technique that uses a new criterion based on the weights of subgroups, enabling the identification of the important features within the dataset. The second stage performs a supervised classification considering the newly learned structure of the a priori classes, the clustering results, and the selected features. The effectiveness of the proposed strategy had been validated also on synthetic data. The results obtained from these analyses showed an improvement in the accuracy of the proposed strategy compared to competitive methods.

CONTRIBUTED SESSION 1 (10:50 – 12:20, JULY 13)
ADVANCES AND APPLICATIONS IN STATISTICAL MODELLING I
Chair: Daniel Jeske (University of California, Riverside)
Room: GSB 308

Improved Semi-Parametric Inference for a Semi-Supervised Mixture Model

Daniel Jeske (University of California, Riverside)

A mixture of a distribution of responses from untreated patients and a shift of that distribution is a useful model for the responses from a group of treated patients. The mixture model accounts for the fact that not all the patients in the treated group will respond to the treatment and consequently their responses follow the same distribution as the responses from untreated patients. The treatment effect in this context consists of both the fraction of the treated patients that are responders and the magnitude of the shift in the distribution for the responders. In this paper, we introduce inference based on a pseudo-likelihood approach and compare it with an existing method of moment approach. An extensive simulation study is used to assess robust performance of the two approaches regarding point estimation, confidence intervals, and confidence regions. The methods are illustrated on an example blood pressure data set. This is joint work with Bradley Lubich.

Unsupervised estimation of loss distributions via dynamic mixtures

Marco Bee (University of Trento)

Maximum likelihood estimation of mixture distributions with dynamic weights is difficult, mostly because of the need to evaluate an intractable normalizing constant. Simulation-based estimation methods are an appealing alternative. We first employ approximate maximum likelihood estimation (AMLE), which is a general method possibly applied to any model, as long as simulation is feasible. The focus is on the dynamic lognormal-generalized Pareto distribution, and the Cramér-von Mises

distance is used to measure the discrepancy between observed and simulated samples. A hybrid procedure is developed, where standard maximum likelihood is first employed to determine the bounds of the uniform priors. Next, we implement a noisy cross-entropy (CE) approach, which also avoids exact evaluation of the normalizing constant. CE is comparable to AMLE in terms of statistical efficiency but is less demanding from the computational point of view. Simulation experiments and a real-data application suggest that both approaches yield a major improvement with respect to standard maximum likelihood estimation. The R package FitDynMix, available at https://github.com/marco-bee/FitDynMix, contains the functions needed for the implementation of MLE, AMLE and noisy CE.

Bias of variable selection for Lasso-based methods: the distributions of predictors matter
Hong Gu (Dalhousie University)

Variable selection based on Lasso shrinkage has been widely used and well developed. In regression it has become an indispensable tool in high dimensional settings when the number of variables is much larger than the number of observations. There is a large body of literature on variable selection methods based on the Lasso penalty and its extensions. However, the theory and simulation studies supporting the use of these methods were mostly developed explicitly or implicitly for multivariate Gaussian predictors, and there is currently no literature assessing the influence of the predictors' distributions on variable selection effects. Standardisation of the predictors for Lasso is recommended as a default to ensure Lasso is scale-invariant. While standardisation in terms of standard deviation is appropriate for normal predictors, the standard deviation is not always such a good measure of scale for heavy-tailed distributions. The lack of a more appropriate standardisation method for heavy-tailed predictors leads to worse performance of Lasso-based methods. In real applications, heavy-tailed predictors are very common. In this talk we present a comprehensive simulation study to evaluate the effects of the predictors' distributions on Lasso-based variable selection methods for generalized linear models and compare their corresponding prediction accuracy. The simulation results show that the predictors' distributions usually have limited effect on variable selection for Gaussian regression models. In contrast, heavy-tailed variables are usually under-selected in Binomial logistic regression, and over-selected in Poisson regression models. Furthermore, this bias in variable selection reduces prediction accuracy in these cases. Box-Cox transformation of the predictors can improve variable selection of some methods, even when it results in mis-specified models, but does not completely remove the impact of predictor distribution.

Structure Selection for Neural Networks
Toby Kenney (Dalhousie University)

A challenge in applying neural networks is selecting the features and structure. Even for simple fully connected feed-forward structures, there are a lot of possibilities for the number of layers and the number of nodes in each layer. Often there are further structures that can account for the relation between predictors - for example, separately fitting multiple groups of predictors and then combining the results. The difficulty in selecting structures has led to these structures being neglected. Another issue is that if too many predictors are included, the network may be subject to overfitting. It is therefore desirable to select only the most useful predictors for the network. There has been previous work on feature selection, and more limited work on structure selection for neural networks. In this talk, I will present a novel method for selecting both the structure and the features of a neural network, to result in an effective sparse neural network structure, which can improve both prediction and interpretation. The basic idea of our method is to start with a dense neural network structure that contains all structures under consideration, then to remove links that do not improve the predictive performance of the network, based on LASSO and backward selection.

CONTRIBUTED SESSION 2 (13:30 – 15:00, JULY 13)
ADVANCES AND APPLICATIONS IN STATISTICAL MODELLING II
Chair: Paulo Rodrigues (Federal University of Bahia)
Room: GSB 308

Anomaly detection, classification, and stock price prediction using Random Forest Algorithm and LSTM model

Leonard Mushunje (Columbia University)

This paper explores whether there are anomalies in high-frequency stock trades in South Africa. Using the JSE daily data from 2010 to 2022, we hypothesize that there are complexities associated with high-frequency stock data, which carries hidden important information, and this information can be helpful to investors. Under high-frequency trading settings, traders should be able to detach and disseminate information from complex sets quickly, efficiently, and profitably. Given any stock portfolio, they should be able to separate the risky stocks from the less risky and rich ones before making any investment decisions. However, this is less attainable in emerging and less technical economies like South Africa, which still rely on traditional trading norms (managerial expertise and emotional trading). Therefore, this paper aims to provide a powerful solution to this fundamental problem. Firstly, we study the time-stamped behavior of stock prices using Long short term memory model (LSTM). We note that JSE stock prices are non-stationary, have fat tails, and have a long memory, which exhibits the ARCH effects and the volatility traits of the stocks. Secondly, we employ the Random Forest algorithm to capture useful stock features further and classify the data quickly. We trained the model hourly to capture the anomaly data, classify trades, and convert them to profitable trades. From this model, we managed to classify stock trades into three categories that are high premium (less risky), premium(satisfactory), and doubtful (high risk). Ideally, volatile stocks with low returns are riskier (doubtful), and true otherwise. We evaluate our RF model using OOB error and cross-validation. Minor prediction errors were reported with an increase in the number of trees, signaling its robustness in capturing the embedded stylized facts about stock trades.

Measurement system analysis for multivariate and functional data

Banefsheh Lashkari (University of Waterloo)

A measurement system analysis is the process of understanding and quantifying the variability in measurement data related to the measurement system. The primary objective of a measurement system analysis is to determine whether the measurement system is suitable for its intended purpose. In our study, we demonstrate the main parameters used in a measurement system study and extend this analysis to measurement data with multivariate or functional forms. We investigate the procedure for estimating the covariance structure of the variation components and explore several options for summarizing the variability structure using a single scalar, along with their corresponding statistical properties. We illustrate the method through a simulation study.

The role of credit risk in the relationship between firm size and efficiency: evidence from Italy

Agnese Rapposelli (University G.d'Annunzio of Chieti-Pescara)

The firm size-efficiency relationship is a topic that is debated at both theoretical and empirical levels. The relevance of this issue is even greater in countries where the production system is characterized by the prevalence of small enterprises, as in the case of Italy, where small and medium-sized enterprises (SMEs) account for 5.3 million active companies, almost 99% of all companies in the country. In this context, the study investigates the relationship between firm size and technical efficiency in Italian manufacturing and service firms under the credit risk lens. Specifically, we use the Data Envelopment Analysis (DEA) technique to measure the efficiency of a sample of large, medium-sized, and small private Italian firms (over 10,000 firms), by employing financial and economic ratios as well as a default measure, namely days past due. To the best of our knowledge this factor, included as an undesirable output in the DEA models considered, has not yet been employed in previous works, allowing us to study the size-efficiency relationship in the context of the bank-firm relationship. Our results confirm the positive relationship between size and efficiency: larger companies are more efficient across all profiles investigated (capital strength, economic-financial performance and relationship with lending banks). Our study demonstrates the need to improve the efficiency of the Italian entrepreneurial system,

consisting mainly of small companies, through their dimensional growth. Hence, our research can provide supervisors and policymakers with useful indications for the allocation of capital to the productive system. On the one hand, business mergers should be encouraged; on the other hand, the allocation of financial resources to small enterprises should be aimed at improving efficiency. This is joint work with Giuliana Birindelli and Michele Modina.

Generalized log-logistic proportional hazard model: a non-penalty shrinkage approach
Shakhawat Hossain (University of Winnipeg)

We consider the pretest and shrinkage estimation methods for estimating regression parameters of the generalized log-logistic proportional hazard model. This model is a simple extension of the log-logistic model, which is closed under the proportional hazard relationship. The generalized log-logistic proportional hazard model also has attributes similar to that of the Weibull model. We consider this model for right-censored data when some of the parameters shrink to a restricted subspace. This subspace information on the parameters is used to shrink the unrestricted model estimates toward the restricted model estimates. We then optimally combine the unrestricted and restricted estimates in order to define pretest and shrinkage estimators. Although this estimation procedure may increase the bias, it also reduces the overall mean squared error. The efficacy of the proposed model and estimation techniques are shown using a simulation study as well as an application to real data. The shrinkage estimator poses less risk than the maximum likelihood estimator when the shrinkage dimension exceeds two; this is shown through simulation and real data application.

CONTRIBUTED SESSION 3 (15:20 – 16:50, JULY 13)
ADVANCES IN THE DESIGN AND ANALYIS OF CLINICAL STUDIES
Chair: Nathaniel Stevens (University of Waterloo)
Room: GSB 308

Fast Sample Size Determination for Two-Group Equivalence Tests with Unequal Variances
Luke Hagar (University of Waterloo)

Two-group equivalence tests are broadly used to show that two quantities are practically equivalent in pharmaceutical, product development, and quality control settings. These equivalence tests are typically conducted using two one-sided t-tests, where the relevant test statistics jointly follow a bivariate t-distribution with singular covariance matrix. Unless we assume the two groups of data have the same variance, the degrees of freedom for this bivariate t-distribution are non-integer and unknown a priori. This makes it difficult to analytically find sample sizes that yield desired statistical power for the equivalence test. As such, functionality to compare two groups with unequal variances is not incorporated in popular R software for equivalence testing design. We propose a novel simulation-based method that uses low-dimensional randomized Sobol' sequences to recommend sample sizes for two-group equivalence tests with unequal variances. This method consistently estimates statistical power much more efficiently than traditional simulation-based approaches. We also extend this method to estimate the power curve using root-finding algorithms. Moreover, our method for sample size determination is widely applicable to two-group equivalence tests facilitated via parallel, crossover, and sequential designs.

Use of Markov decision processes for statistical design of response adaptive clinical trials
Xikui Wang (University of Manitoba)

Clinical trials are regarded as the most reliable and efficient way to evaluate the efficacy of new medical interventions. This practice has taken a prominent role in the pharmaceutical industry. However, clinical experimentation on human subjects requires a careful balancing act between the benefits of the collective (i.e., exploration) and the benefits of the individual (i.e., exploitation). Response adaptive designs represent a major advancement in clinical trial methodology that helps balance these ethical issues and improve efficiency without undermining the validity and integrity of the clinical research. Under suitable statistical approaches, the response adaptive randomization process can be formulated as a Markov decision process. Based on our recent research (jointly with Yanqing Yi), we discuss how the use of Markov decision processes improves ethics of the clinical research while safeguarding the quality of statistical evaluation of the clinical trial.

Nonparametric estimation of the distribution of latent treatment effects within the context of a control group versus treatment group design

Benjamin Ellis (University of California, Riverside)

Control versus treatment clinical trials are prone to averaging over individual effects to produce a one size fits all conclusion: does a drug/procedure work, on average? However, it is often the case that subjects have their own unique response to treatment. We have developed an infinite mixture model that can produce a nonparametric estimate of the latent distribution of treatment effects for a population. The estimated distribution of treatment effects enables researchers to make conclusions about what proportion of the population will have a treatment effect of a specified size (i.e., medically significant effect), without assuming a specific distribution on their data.