5th Waterloo Student Conference in Statistics, Actuarial Science and Finance

Book of Abstracts

Day 1

Keynote (Chair: Augustine Wigle)

Name: Aaditya Ramdas

Title: A game-theoretic theory of statistical evidence

Abstract: This talk will describe an approach towards testing hypotheses and estimating functionals that is based on games. In short, to test a (possibly composite, nonparametric) hypothesis, we set up a game in which no betting strategy can make money under the null (the wealth is an "e-process" under the null). But if the null is false, then smart betting strategies will have exponentially increasing wealth. Thus, hypotheses are rewritten as constraints in games, the statistician is a gambler, test statistics are betting strategies, and the wealth obtained is directly a measure of evidence which is valid at any data-dependent stopping time (an e-value). The optimal betting strategies are typically Bayesian, but the guarantees are frequentist. This "game perspective" provides new statistically and computationally efficient solutions to many modern problems, like nonparametric independence or two-sample testing by betting, estimating means of bounded random variables, testing exchangeability, and so forth. The talk will summarize work from many papers in the literature.

Session 1A: Biostatistics (Chair: Grace Tompkins)

Talk 1

Name: Liyuan Zheng

Title: Control the false discovery rate with complex auxiliary information

Abstract: In many genetic multiple testing applications, the signs of test statistics provide critical directional insights. A negative sign suggests that the expression of the corresponding gene is potentially suppressed, while a positive sign indicates a potentially elevated expression level. However, most existing procedures that control the false

discovery rate (FDR) ignore such valuable information. In this paper, we propose the covariate and direction adaptive knockoff (Codak) procedure, to incorporate additional auxiliary information, including directional and covariate information. We demonstrate that the Codak procedure effectively controls the FDR in finite samples. Through simulation studies and real data analysis, we show that our procedure is competitive to existing covariate-adaptive methods. The companion R package Codak is available.

Talk 2

Name: Henan Xu

Title: Gender Disparities in Rehospitalisations after Coronary Artery Bypass Grafting: Evidence from a Functional Causal Mediation Analysis of the MIMIC-IV data

Abstract: Hospital readmissions following coronary artery bypass grafting (CABG) not only impose a substantial cost burden on healthcare systems but also serve as a potential indicator of the quality of medical care. Previous studies of gender effects on complications after CABG surgery have consistently revealed that women tend to suffer worse outcomes. To better understand the causal pathway from gender to the number of rehospitalisations, we study the postoperative central venous pressure (CVP), frequently recorded over patients' intensive care unit (ICU) stay after the CABG surgery, as the mediator. Confronted with time-varying CVP measurements and zero-inflated rehospitalisation counts within 60 days following discharge, we propose a parameter-simulating quasi-Bayesian Monte Carlo approximation method that accommodates a functional mediator and a zero-inflated count outcome for causal mediation analysis. We find a causal relationship between the female gender and increased rehospitalisation counts after CABG, and that time-varying central venous pressure mediates this causal effect.

Talk 3

Name: Olana Dabi

Title: Diagnostic test accuracy meta-analysis based on exact within-study variance estimation method

Abstract: Meta-analysis of diagnostic test accuracy studies commonly synthesizes study-specific test sensitivity and test specificity from different studies that aim to quantify the screening or diagnostic performance of a common index test of interest. A bivariate

random effects model that utilizes the logit transformation of sensitivity and specificity and accounts for the within- and between-study heterogeneity is commonly used to make statistical inferences about the unknown test characteristics. However, it is well reported that this model may lead to misleading inference since it employs the logit transformation and approximate within-study variance estimate. Alternative transformations which do not require continuity corrections, such as the arcsine square root and Freeman-Tukey double arcsine, were recently proposed to overcome the former limitation. However, these solutions also suffer from using approximate within-study variance estimates, which can only be justified when within-study sample sizes are large. To overcome these problems, an exact within-study variance estimation approach is proposed, which does not require a continuity correction and is invariant to transformations. The new method and the existing approaches are evaluated using real-life and simulated meta-analytic data.

Talk 4

Name: Diribsa Bedada

Title: An evaluation of the robustness of the generalized linear mixed model for an individual participant data meta-analysis of diagnostic test accuracy studies

Abstract: Diagnostic tests are essential in clinical practice to identify patients with a particular disease or condition. The accuracy of these tests is typically evaluated through sensitivity (Se) and specificity (Sp), which measure a test's ability to correctly identify diseased and non-diseased individuals, respectively. Inaccurate test results can lead to harmful treatment decisions, making the reliability of these tests crucial. Meta-analysis, commonly used to pool data from multiple diagnostic test accuracy studies (DTAs), frequently applies a generalized linear mixed model (GLMM) to account for variability across studies. However, this model often assumes that random effects follow a bivariate normal distribution, which may not hold in practice and may lead to misleading results if the assumption is violated. Thus, this study aims to investigate the robustness of the standard GLMM for an individual participant data meta-analysis (IPDMA) of DTAs by considering alternative random effects distributions, specifically the bivariate skew-normal and bivariate t-distributions, to evaluate the model's robustness when data follow asymmetric and heavier-tailed distributions. Adaptive Gauss-Hermite quadrature is used to estimate both the fixed and random effects parameters. Through extensive simulations, this study will evaluate the impact of distributional assumptions on the estimation of Se and Sp as well as the between-study variances and covariance. Absolute bias, root mean squared error, and coverage probability will be among the performance measures used to

appraise the model's performance when its distributional assumption is not met. The findings of this simulation study will be used to inform future methodological developments to overcome the limitations of the standard GLMM for an IPDMA of DTAs.

Session 1B: Risk Measures and Risk Management (Chair: James Huang) (10:45)

Talk 1

Name: Qinghua Ren

Title: Counter-monotonic risk allocations and distortion risk measures

Abstract: In this paper, we provide a systematic study of efficient risk sharing in markets where allocations are constrained to be counter-monotonic. The preferences of the agents are modelled by a common distortion risk measure, or equivalently, by a common Yaari dual utility. We consider three different settings: risk-averse agents, risk-seeking agents, and those with an inverse S-shaped distortion function. In each case, we provide useful characterizations of optimal allocations, for both the counter-monotonic market and the unconstrained market. To illustrate our results, we consider an application to a portfolio choice problem for a portfolio manager tasked with managing the investments of a group of clients, with varying levels of risk aversion or risk seeking. We determine explicitly the optimal investment strategies in this case. Our results confirm the intuition that a manager investing on behalf of risk-seeking agents tends to invest more in risky assets than a manager acting on behalf of risk-averse agents.

Talk 2

Name: James Huang

Title: g-monotonicity, additivity and risk aversion

Abstract: We introduce a collection of dependence structures weaker than comonotonicity called g-comonotonicity, determined by an increasing function g on [0, 1]. A subclass of such functions also generate a family of risk measures called spectral risk measures. This paper develops a connection between the two objects, related to the question: when does worst case aggregation, in this case additive aggregation, occur?

Talk 3

Name: Alessandro Mutti

Title: Convex bounds on sums with generalized FGM copula

Abstract: Building on the one-to-one relationship between generalized FGM copulas and multivariate Bernoulli distributions, we prove that the class of multivariate distributions with generalized FGM copulas is a convex polytope. Therefore, we find sharp bounds in this class for many aggregate risk measures, such as value-at-risk, expected shortfall, and entropic risk measure, by enumerating their values on the extremal points of the convex polytope. This is infeasible in high dimensions. We overcome this limitation by considering the aggregation of identically distributed risks with generalized FGM copula specified by a common parameter p. In this case, the analogy with the geometrical structure of the class of Bernoulli distribution allows us to provide sharp analytical bounds for convex risk measures.

Talk 4

Name: Yimiao Zhao

Title: Catastrophe Risk Pooling

Abstract: The ongoing climate change causes natural disaster losses are on the rise. The huge gap between actual and insured losses, however, is growing. Insurance-based solutions are receiving increasing attention to mitigate disaster and climate risks. Due to the features of catastrophic risks, the catastrophe insurance market has failed to provide sufficient protection. In this paper, we discuss how pooling catastrophe risks from various geographies and perils can be an effective tool for diversifying catastrophic risks and reducing risk premiums. Simulations and real data analysis are provided to illustrate our results.

Session 2A: Theoretical Statistics (Chair: Madison Cranstone)

Talk 1

Name: Zhaoqi Yang

Title: On Retrieving Bivariate Data Sets from Their Moments

Abstract: This presentation introduces several methodologies that solve the inverse problem of recovering a bivariate sample from subsets of its associated marginal and joint integer moments or the ranks of the observations within each component. These results rely in part on their univariate counterpart, which is examined in some detail. It is also explained that they also apply to complex-valued data sets, and can be extended to multivariate observations. Several illustrative examples are presented.

Talk 2

Name: Pengqi Liu

Title: Regularization in finite mixture of sparse GLMs with ultra-high dimensionality

Abstract: Finite mixture of generalized linear regression models (FM-GLM) are used for analyzing data that arise from populations with unobserved heterogeneity. In recent applications of FM-GLM, data are often collected on a large number of features. However, fitting an FM-GLM to such high-dimensional data is numerically challenging. To cope with the high-dimensionality in estimation, it is often assumed that the model is sparse and only a handful of features are relevant to the analysis. Most of the existing development on sparse estimation is in the context of homogeneous regression or supervised learning problems. In this talk, I will discuss some of the challenges and recent computational and theoretical developments for sparse estimation in FM-GLM when the number of features can be in exponential order of the sample size. I will also discuss our results in finite sample settings via simulations.

Talk 3

Name: Alysha Cooper

Title: Dominating hyperplane regularization for sparse multivariate count regression

Abstract: Multivariate count data, measured as taxa counts, are common in many fields such as genetics, microbiology, and ecology. Identifying the factors that influence these counts is an important biological question. While multinomial regression is a common approach for modeling these data, variability in taxa counts is often higher than what is assumed by the multinomial model. The Dirichlet-multinomial (DM) regression model addresses this overdispersion, but estimating regression coefficients becomes challenging, especially as the number of potentially relevant factors increases. Variable selection becomes necessary to handle this complexity, and the sparse group lasso (SGL)

offers a natural solution for DM regression models. In this talk, I will introduce the novel Dominating Hyperplane Regularization via the majorization-minimization (MM) framework. DHR provides a stable and simple optimization algorithm for estimation of regularized DM regression with the SGL penalty. I will show how DHR applied to the SGL penalty gives rise to a surrogate function that can be expressed as a weighted ridge penalty and consequently, we can use an iteratively re-weighted Poisson ridge regression to estimate regularized DM with the SGL penalty. Superior model performance is demonstrated through simulation and an application to benthic compositions from the Alberta oil sands region.

Talk 4

Name: Qinyu Wu

Title: On Generalization and Regularization via Wasserstein Distributionally Robust Optimization

Abstract: Wasserstein distributionally robust optimization (DRO) has found success in operations research and machine learning applications as a powerful means to obtain solutions with favourable out-of-sample performances. Two compelling explanations for the success are the generalization bounds derived from Wasserstein DRO and the equivalency between Wasserstein DRO and the regularization scheme commonly applied in machine learning. Existing results on generalization bounds and the equivalency to regularization are largely limited to the setting where the Wasserstein ball is of a certain type and the decision criterion takes certain forms of an expected function. In this paper, we demonstrate that generalization bounds and equivalence to regularization can be achieved in a significantly broader setting, where the Wasserstein ball may be of a general type, and the decision criterion can be a general measure of risk, i.e., nonlinear in distributions. This not only accommodates important machine learning and operations management applications, but also expands to general decision-theoretical frameworks previously unaddressed using Wasserstein DRO. Our results are strong in that the generalization bounds do not suffer from the curse of dimensionality and the equivalency to regularization is exact. As a byproduct, our regularization results broaden considerably the class of Wasserstein DRO models that can be solved efficiently via regularization formulations. These general assurances should serve as a strong basis for expanding the application of Wasserstein DRO across diverse domains of data-driven decision problems.

Session 2B: Actuarial Science: Theory and Application (Chair: Yixin Ding)

Talk 1

Name: Benjamin Côté

Title: Centrality and topology properties in a tree-structured Markov random field

Abstract: The topology of the tree underlying a tree-structured Markov random field (MRF) is central to the understanding of its stochastic dynamics: it is, after all, what synthesizes the rich dependence relations within the MRF. The aim of this talk is to shed light on the influence of the tree's topology, through an extensive comparison-based analysis, on the aggregate distribution of the MRF. This is done within the framework of a recently introduced family of tree-structured MRFs with the uncommon property of having fixed marginal distributions unaffected by the dependence scheme. We establish convex orderings of sums of MRFs encrypted on trees having different topologies, leading to the devising of a new poset of trees. We present an exhaustive investigation of this new poset. Such an analysis requires, beforehand, to study the joint distribution of a

MRF's component and its sum, a dependence relation we refer to as aggregate influence. We employ dependence orders to compare aggregate influences within a tree according to the position of their associated vertices. The resulting orderings are reflected through allocation-related quantities, which thus act as centrality indices. This motivates a discussion on network centrality, a notion stemming from the study of social networks, in the context of MRFs.

Talk 2

Name: André Orelien Chuisseu Tchuisseu

Title: Actuarial fire-spreading model based on tree-structured Markov Random Fields, with insurance applications

Abstract: We present an actuarial fire-spreading model based on tree-structured Markov Random Fields (MRF) with applications to property fire insurance. We propose a general framework to model fire propagation, where each site is structured as a tree with units connected by stochastic dependencies. Our analysis focuses on tree-structured sites with four and five units. For each configuration, we provide a stochastic representation of the total fire losses. We explain how to find the distribution of the total fire losses depending on the configuration. We analyze the impact of propagation probabilities on the distribution of total fire losses using stochastic orders. The investigation compares five

tree structures by employing the stop-loss order to highlight the differences in risk profiles between structures. The model is further examined through numerical applications offering insights into the practical implications for insurance, particularly in calculating premiums and risk measures. This approach provides a useful tool for assessing fire risk in complex building structures, with analytical advantages over simulation-based methods.

Talk 3

Name: Emma Kroell

Title: Model Ambiguity in Risk Sharing with Monotone Mean-Variance Criteria

Abstract: We consider the problem of an agent who faces losses over a finite time horizon and may choose to share some of these losses with a counterparty. The agent is uncertain about the true loss distribution and has multiple models for the losses. Their goal is to optimize a mean-variance type criterion with model ambiguity through risk sharing. We construct such a criterion by adapting the monotone mean-variance preferences of Maccheroni et al. (2009) to the multiple models setting, exploiting a dual representation to mitigate time consistency issues. Assuming a Cramer-Lundberg loss model, we explicitly obtain the optimal risk sharing contract and the agent's wealth process under the optimal strategy. We apply the optimal strategy to an insurance setting using data from two different Spanish automobile insurance portfolios and illustrate the results.

Talk 4

Name: Sebastian Calcetero Vanegas

Title: A population sampling framework for claim reserving in general insurance

Abstract: Claim reserving in insurance has been explored within two major frameworks: the macrolevel approach, which estimates reserves at an aggregate level (e.g., Chain-Ladder), and the micro-level approach, which estimates reserves individually (e.g Antonio, K., & Plat, R. (2014)). These approaches, rooted in entirely different theoretical foundations, are somewhat incompatible, hindering practitioners from adopting more flexible models. In this talk, we introduce a novel unified statistical framework for reserving methods approach grounded in population sampling theory. We demonstrate that macro and micro-level models represent extreme yet natural instances of an augmented inverse probability weighting (AIPW) estimator, enabling a seamless transition by integrating principles from both aggregate and individual models simultaneously into more accurate

estimations. Additionally, we go further and explore how other population sampling motivated tools, not explored in reserving, can be employed to enhance the robustness and effectiveness of current reserving models.

Session 3: Machine learning (Chair: Bryn Crandles)

Talk 1

Name: Tiancheng Yang

Title: exKidneyBERT: a language model for kidney transplant pathology reports and the crucial role of extended vocabulaires

Abstract: Background. Pathology reports contain key information about the patient's diagnosis as well as important gross and microscopic findings. These information-rich clinical reports offer an invaluable resource for clinical studies, but data extraction and analysis from such unstructured texts is often manual and tedious. While neural information retrieval systems (typically implemented as deep learning methods for natural language processing) are automatic and flexible, they typically require a large domain-specific text corpus for training, making them infeasible for many medical subdomains. Thus, an automated data extraction method for pathology reports that does not require a large training corpus would be of significant value and utility.

Objective. to develop a language model-based neural information retrieval system that can be trained on small datasets and validate it by training it on renal transplant-pathology reports to extract relevant information for two predefined questions.

Methods. The study aimed to develop a neural information retrieval system that can be successfully trained on small text corpuses. We develop such a system and validate it by training it to automatically answer two pre-defined questions given text from renal pathology reports: 1) "What kind of rejection does the patient show?"; and 2) "What is the grade of interstitial fibrosis and tubular atrophy (IFTA)?". First, we followed the conventionally recommended procedure for developing domain-specific models and pre trained a previously proposed medical language model called Clinical BERT further with our text corpus, which contains 3.4K renal transplant pathology reports and 1.5M words, using Masked Language Modeling to obtain 'Kidney BERT'. Second, we hypothesized that the conventional pre-training procedure fails to capture the intricate vocabulary of narrow technical domains. We created extended Kidney BERT ('exKidneyBERT') by extending the tokenizer of Clinical BERT with six technical keywords from our corpus (which we determined were missing from the original tokenizer vocabulary) and then repeating the

pre-training procedure. Third, to further improve performance, all three models were fine tuned with information retrieval (IR) heads tailored to the two questions of interest.

Results. For the first question regarding rejection, the overlap ratio at word level for exKidneyBERT – 83.3% for antibody-mediated rejection (ABMR) and 79.2% for T-cell mediated rejection (TCMR) – beats that of both Clinical BERT and Kidney BERT (both are 46.1% for ABMR, and 65.2% for TCMR). For the second question regarding IFTA, the exact match rate of exKidneyBERT (95.8%) beats that of Kidney BERT (95.0%) and Clinical BERT (94.7%).

Conclusion. We developed exKidneyBERT, a high-performing model for automatically extracting information from renal pathology reports. More broadly, we found that when working in domains with highly specialized vocabulary, it is essential to extend the vocabulary library of the BERT tokenizer to improve model performance, otherwise, pre training (especially on small corpuses) is ineffective. In our case, pre-training BERT language models on kidney pathology reports improved model performance even though the training corpus was much smaller than the corpora normally used to train language models.

Talk 2

Name: Julie Wojtiw-Quo

Title: Evaluating air quality using isolation forests - Clearing the air on pollution hotspots

Abstract: Climate change is one of many relevant topics to society that can be studied through statistical methodologies. One important, measured indicator of climate change is air quality. Many indicators of air quality are collected and published openly, including carbon dioxide and carbon monoxide levels, methane levels, and other greenhouse gas levels. This term project will focus on a case study involving such indicators collected across the city of Montreal. Specifically, after preprocessing and cleaning the data, a preliminary analysis and exploration will be conducted on the ACES (Atmospheric Composition Environmental Survey) 2023 Montreal dataset. Then, an isolation forest will be run on the most important variables to identify hotspots within the city of Montreal. These hotspots will be co-referenced with building zoning, use, height, and other variables in order to classify abnormal outdoor air pollution and carbon dioxide levels. Today's talk will focus on an introduction to isolation forests, alternative statistical methods, and the planned methodology for undertaking this term project.

Talk 3

Name: Ambrose Emmett-Iwaniw

Title: Enhancing neural autoregressive distribution estimators with low-discrepancy orderings

Abstract: Given the importance of ordering pixels has on the training of autoregressive models such as NADE, MADE, PixelRNN, and ConvNADE, we present a new method for ordering pixels by utilizing the properties of low-discrepancy point sets from quasi-Monte Carlo (QMC) theory. We detail how to generate so-called low-discrepancy pixel orderings and apply this method to a novel variant of the convolution NADE model employing the use of Beta distribution to model the response variable, rather than the common use of a mixture of Gaussian (MoG). Comparisons are made on the MNIST, FER2013 and LHQ datasets illustrating in all cases using low-discrepancy pixel ordering has lower error loss value on test results and more realistic reconstructed images, a task which is closely related to image completion and inpainting. We also present realistic generated images from the LHQ dataset for different orderings. This work also shows how Neural Autoregressive Distribution Estimators are able to generate natural images.

Talk 4

Name: Gradon Nicholls

Title: Model-assisted double-coding of open-ended survey questions with large language models

Abstract: Open-ended questions allow survey respondents to provide answers in their own words without being constrained or influenced by pre-specified answer options. On the other hand, the textual data resulting from open-ended questions are generally more difficult to analyze than responses from closed-ended questions, and so to facilitate analysis, text responses are often manually coded. Having two independent coders ("doublecoding") has been proposed to improve coding quality, but disagreements among the two coders must then be resolved (e.g. through discussion or a third coder). The objective of the current research is to determine to what extent double-coding can improve coding quality when one of the human coders is replaced by model predictions. Clearly, it is not sufficient for the machine coder to exactly mimic the human coder's behaviour--that is, it must (at least occasionally) catch mistakes made by the human coder. In this context, we investigate the utility of pre-trained Large Language Models (LLMs) combined with text

data augmentation techniques to increase the number of examples seen during model training.

Talk 5

Name: Argho Das

Title: Social Unrest in Bangladesh: A Sentiment and Emotion Analysis of Public Opinion During the 2024 Student Quota Protests

Abstract: This research presents a comprehensive sentiment and emotion analysis of Twitter data during and after the political protests in Bangladesh, leveraging natural language processing techniques and machine learning models such as DistilBERT and RoBERTa. The study focused on public opinion trends surrounding the 2024 student-led protests, analyzing both sentiment (positive, negative, and neutral) and emotions (e.g., anger, joy, fear) across a total of 1182 tweets during the protest and 2500 tweets afterward. A comparison of sentiment distribution before and after the protests reveals a dominant negative sentiment, though positive sentiment showed a slight increase post-protest. This suggests that while dissatisfaction persisted, there was also some optimism in the aftermath. Emotionally, neutral and negative emotions prevailed in both phases, but anger diminished post-protest, and emotions like joy saw an uptick, indicating a shift in the public's emotional response over time. Data augmentation techniques using GPT-2 were employed to expand the dataset, enhancing the reliability of results. This research provides key insights into the role of social media in shaping political discourse and highlights the value of applying NLP to understand social movements and their societal impact. The analysis was visualized through percentage-based sentiment and emotion comparisons for greater clarity and accuracy.

Day 2

Keynote (Chair: Zachary Van Oosten)

Name: Etienne Marceau

Title: Tree-Structured Ising models: Mean parameterization, efficient computation methods and stochastic ordering

Abstract: High-dimensional multivariate Bernoulli distributions are essential in the modeling of binary data in actuarial contexts. Tree-structured Ising models, a class of undirected graphical models for binary data, have been proven to be useful in a variety of applications in machine learning. We assess advantages of expressing tree-based Ising models via their mean parameterization rather than their commonly chosen canonical parameterization. This includes fixed marginal distributions, often convenient for dependence modeling, and the dispelling of the intractable normalizing constant otherwise plaguing Ising models. We derive an analytic expression for the joint probability generating function of mean-parameterized tree-structured Ising models. The latter is used to build efficient computation methods for the sum of its constituent random variables. Similarly, we derive an analytic expression of their ordinary generating function of expected allocations, providing means for exact computations in the context of risk allocations. We examine the dependence properties of the tree-structured Ising models and their applications in actuarial science.

Session 4A: Methods for Inference and Improving Efficiency (Chair: Laura Bumbulis)

Talk 1

Name: Laura Bumbulis

Title: Robustness and Efficiency Considerations when Testing Process Reliability with a Limit of Detection

Abstract: Processes in biotechnology are considered reliable if they produce samples satisfying regulatory benchmarks. For example, laboratories may be required to show that levels of an undesirable analyte rarely (e.g. in less than 5% of samples) exceed a tolerance threshold. This can be challenging when measurement systems feature a lower limit of detection rendering some observations left-censored. In this talk we discuss the implications of detection limits for location-scale model-based inference in reliability

studies, including their impact on large and finite sample properties of various estimators; power of tests for reliability and goodness of fit; and sensitivity of results to model misspecification. To improve robustness we then examine other approaches, including exact binomial tests and a weakly parametric method where the right tail of the response distribution is approximated using a piecewise constant hazard model. Time permitting, we discuss simulations to inform sample size selection in future reliability studies and an application to residual white blood cell levels in transfusable blood products.

Talk 2

Name: Victor Malinowski

Title: Investigating the Performance of Direct and Indirect Causal Effect Estimators under Partial Interference and Structured Nearest Neighbour Interference

Abstract: In the framework of causal inference, interference occurs when one subject's treatment has a causal effect on another subject's potential outcomes. This indirect causal effect has shifted from being viewed as a nuisance in the past to being the primary causal effect of interest in many contexts. This presentation outlines the methods proposed by Tchetgen and VanderWeele (2012) to estimate the population average direct, indirect, total, and overall causal effects and to quantify their uncertainty in data exhibiting stratified partial interference using Hajek style IPW point estimators and sandwich form variance estimators. We then conduct a simulation study demonstrating these estimators consistently and efficiently estimate causal indirect effects not only in stratified partial interference settings, but also in data generated under structured nearest neighbour interference. We then apply the outlined methods and simulation study results to an agronomy dataset where we answer a relevant question from the literature regarding whether one crop's emergence date has a causal effect on another crop's grain yield by simultaneously testing for stratified partial interference and structured nearest neighbour interference.

Talk 3

Name: Kyu Min Shim

Title: Leveraging Counterfactual Estimation for Effective Variance Reduction in Online Controlled Experiments

Abstract: Reducing the variance of the average treatment effect estimator is a critical problem in the context of online controlled experiments. Recent developments in variance reduction utilize pre-experiment data to achieve significant variance reduction under the assumption that pre-experiment and in-experiment data are highly correlated. However, in dynamic settings such as e-commerce and social media where trends in data may change rapidly, the validity of such an assumption may be questionable. This work addresses this challenge with a novel variance reduction technique that frames the problem in terms of counterfactual estimation. Inference is then made through pairwise comparisons of observed outcomes and their counterfactual estimates. This method is proven to be asymptotically unbiased, with asymptotic variance scaling with the model's predictive accuracy.

Talk 4

Name: Matteo Gasparin

Title: Combining exchangeable p-values

Abstract: The problem of combining p-values is an old and fundamental one, and the classic assumption of independence is often violated or unverifiable in many applications. There are many well-known rules that can combine a set of arbitrarily dependent p-values (for the same hypothesis) into a single p-value. We show that essentially all these existing rules can be strictly improved when the p-values are exchangeable, or when external randomization is allowed (or both). For example, we derive randomized and/or exchangeable improvements of well known rules like “twice the median” and “twice the average”, as well as geometric and harmonic means. Exchangeable p-values are often produced one at a time (for example, under repeated tests involving data splitting), and our rules can combine them sequentially as they are produced, stopping when the combined p-values stabilize. Our work also improves rules for combining arbitrarily dependent p values, since the latter becomes exchangeable if they are presented to the analyst in a random order. The main technical advance is to show that all existing combination rules can be obtained by calibrating the p-values to e-values (using an α-dependent calibrator), averaging those e-values, converting to a level-α test using the Markov inequality, and finally obtaining p-values by combining this family of tests; the improvements are delivered via recent randomized and exchangeable variants of Markov’s inequality.

Session 4B: Quantitative Finance (Chair: Alexandra Mossman)

Talk 1

Name: Wei Li Fan

Title: A new type of CEV model. Properties, comparison and application to portfolio optimization

Abstract: We introduce the LVO-CEV model, a new type constant elasticity of volatility (CEV) model in which excess return varies linearly with volatility. The model offers strong and weak solutions, depending on the elasticity parameter, thanks to a connection to radial Ornstein-Uhlenbeck processes. Attainability of lower bounds and the existence of pricing measures are discussed. The model provides closed-form solutions within expected utility theory for investors with hyperbolic absolute risk aversion (HARA) utilities on terminal wealth and consumption. Empirical estimation and implementation on indexes and stocks, along with other models, facilitate a comprehensive portfolio management comparison. Our results indicate that the LVO-CEV and M-CEV models outperform the other two models (GBM and CEV). Furthermore, the LVO-CEV model demonstrates reduced rebalancing variability, enhancing its practicality compared to the embedded GBM model. Sensitivity analyses are conducted to explore the impact of parameters on optimal strategy and consumption within the LVO-CEV framework.

Talk 2

Name: Chenxin Lyu

Title: Numerical Fourier method and Multi-dimensional Second-order Taylor scheme for Stochastic Differential Equations

Abstract: In asset pricing, traditional models like the Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) have significantly influenced the field. However, they often fall short in capturing the complexities of modern markets. In this presentation, I will discuss my thesis, which focuses on enhancing asset pricing methodologies through advanced mathematical tools. We begin with a historical overview, tracing the evolution of economic theories and key models such as the Black-Scholes model and the Feynman-Kac formula. I will then probe into the construction and application of stochastic differential equations, Brownian motion, and Itô's stochastic integrals—foundational concepts in financial mathematics. A key highlight of my work is the introduction of an innovative second-order discretization scheme derived from the Fourier expansion of the Brownian bridge process. I will present our accurate approximations of the characteristic function, essential for solving forward stochastic differential equations. Additionally, I will

cover a backward discretization scheme that utilizes the Fourier method for estimating function values over time, showcasing its application in pricing European options through numerical experiments that demonstrate its accuracy. Overall, this work contributes to asset pricing by developing new methods that address the limitations of existing models, providing both theoretical insights and practical tools for financial analysts and researchers.

Talk 3

Name: Xize Ye

Title: Affine Generalized Autoregressive Conditionally Stochastic Heteroskedasticity (GARCSH): Motivation and Applications

Abstract: Typical GARCH models are proven successful in capturing time-varying conditional variance of asset return. Nonetheless, the construction that only 1 innovation drives both the return and variance process makes it difficult to reconcile historical return and forward-looking information, such as the VIX (volatility index). Instead, we directly extend the traditional GARCH models to incorporate stochastic volatility in the modelling by adding an additional innovation to the model, hence resulting in a 2-shock model named Generalized Autoregressive Conditionally Stochastic Heteroskedasticity (GARCSH). In this talk, we give a detailed introduction to its motivation and implementation. We also provide numerical results demonstrating the financial implications as well as areas for future research.

Talk 4

Name: Xipeng Huang

Title: Would a two-benchmark regime be better?

Abstract: After the manipulation scandal during the global financial crisis, LIBOR was phased out, and regu- lators around the world have since transitioned to risk-free reference rates (RFRs). In the United States, the Secured Overnight Financing Rate (SOFR), which is an RFR based on overnight repo transactions, has been designated as the sole replacement for LIBOR. Meanwhile, Europe and Japan have chosen to establish a two-benchmark regime consisting of an RFR and a LIBOR-like credit-sensitive reference rate (CSR). In this study, we consider a two-agent model with one representative firm and one repre- sentative bank in the economy, and use a tractable model to compare the single

RFR regime against the two-benchmark regime with an RFR and a CSR. It turns out that introducing a new credit-sensitive benchmark in addition to the existing risk-free benchmark always improves total welfare but not neces- sarily for both the bank and the firm in some scenarios. Our study also indicates that credit supply is higher and borrowing cost is lower in the two-benchmark system than would be the case for the single RFR system. Furthermore, RFR could be driven out of the market when the CSR and the bank's funding cost have a strong enough correlation. The impacts from fixed rate lending and interest rate swap trading are also discussed.

Session 5A: Computational Statistics and Change Point Detection (Chair: Sijie Chen)

Talk 1

Name: Mohan Wu

Title: Variational Inference for Stochastic Differential Equations

Abstract: Stochastic differential equations (SDEs) are ubiquitous in the natural sciences, engineering, and finance. Due to the intractability of the likelihood function, parameter inference for SDEs typically involves introducing a large number of latent variables that must be integrated out numerically. Traditionally, this is done using Markov chain Monte Carlo (MCMC) methods based on particle filters. However, such MCMC algorithms are prohibitively slow for all but very small problems. In this work, we propose a method of variational inference (VI) for SDE parameter estimation. Our variational model efficiently captures the structure and information propagation in the true posterior distribution using recurrent neural networks (RNNs). Since VI converts integration into an optimization problem and RNNs are highly amenable to efficient computation on modern architectures involving graphical processing units (GPUs), our VI method offers several orders of magnitude speedup over particle filter MCMC. Several examples demonstrate the effectiveness of this approach.

Talk 2

Name: Mingwei Xu

Title: A Bayesian collocation integral method for system identification of ordinary differential equations

Abstract: The ordinary differential equations (ODEs) are widely considered to model the dynamics of complex systems across various scientific areas. To identify the structure of high-dimensional sparse ODEs from noisy time-course data, most existing methods adopt a frequentist perspective, while uncertainty quantification in parameter estimation remains challenging. Under the additive ODE model assumption, we present a Bayesian hierarchical collocation method to better quantify the uncertainty. Our framework unifies the likelihood, integrated ODE constraints and a group-wise sparse penalty, allowing for simultaneous system identification and trajectory estimation. We demonstrate the superior performance of the purposed method through simulation studies, where the recovered system trajectories and estimated additive components are compared with other recent methods. A real data example is also provided.

Talk 3

Name: Jeremy VanderDoes

Title: Change point analysis for functional data using empirical characteristic functionals

Abstract: We develop a new method to detect change points in the distribution of functional data based on integrated CUSUM processes of empirical characteristic functionals. Asymptotic results are presented under mild conditions allowing for low-order moments and serial dependence in the data establishing the limiting null-distribution of the proposed test statistics, as well as their consistency to detect and localize change points in the distribution of functional data. A key consideration in defining these test statistics is the measure used to integrate the CUSUM process over function space. We show that using a measure generated by Brownian motion leads to generally consistent tests. Further, using this measure allows for computationally simple approximations of the necessary integrals, as well as simulation and permutation-based methods to calibrate detection thresholds for change point analysis. The proposed methods are thoroughly investigated and compared to other existing functional data change point methods in simulation experiments, and are further demonstrated by applications to detect change points in models for continuous electricity demand and high-frequency asset price returns.

Talk 4

Name: Ahmed Mozaffari

Title: Non-parametric multiple change point detection for manifold data

Abstract: In this presentation, non-parametric change point detection for stochastic processes defined on manifold spaces M is proposed. The proposed statistic is based on the rank of depths to detect changes in distribution. Theoretical investigation is carried out to obtain finite-sample and asymptotic convergence bounds for the location and detection estimators based on the geometrical properties of the manifold spaces. Finite-sample Monte-Carlo simulation studies endorses the validity of the obtained convergence bounds as well as the detection power of the test statistic in the presence of extreme data and outliers.

Session 5B: Portfolio Management (Chair: Henan Xu)

Talk 1

Name: Zhiqiao Song

Title: Worst-case downside risk measures under a reward-penalty framework with applications to robust portfolio selection

Abstract: In this talk, we present a portfolio selection problem utilizing a reward-penalty framework under distributional uncertainty. We assume the joint distribution of the underlying asset losses is uncertain but falls within a multivariate uncertainty set. Our goal is to find the optimal portfolio by minimizing the worst-case values of downside risk measures. We begin by deriving the worst-case values with the classic mean-variance uncertainty set under these measures and obtaining explicit closed-form solutions for the optimal portfolios. We then extend this classic mean-variance set to a generalized uncertainty set that provides additional flexibility. This set allows the mean of the underlying asset losses to lie within an ellipsoid around the inferred mean, while bounding the discrepancy between the covariance of the underlying asset losses and the inferred covariance. Under this generalized uncertainty set, we reform the robust portfolio selection problems into convex optimization problems. Moreover, we apply real market data to highlight the practical benefits of the reward-penalty structure, comparing the resulting portfolio with those from the classic models and discussing the impact of model parameters on the portfolio performance.

Talk 2

Name: Hamidreza Masoumi

Title: Outperforming Buy-and-Hold: A market Timing Model Using Analysts' Forecasts

Abstract: This paper introduces a market timing model that integrates price trend indicators with analysts' forecasts. The findings demonstrate that incorporating analysts' forecasts significantly enhances market timing decisions. Our model underscores the predictive power of these forecasts for stock returns. By applying this model, both average and risk-adjusted returns outperform a simple buy-and-hold strategy, showcasing its effectiveness in active trading. Additionally, we perform an extensive out-of-sample investment analysis to evaluate the contribution of analysts' forecasts.

Talk 3

Name: Agassi Iu

Title: A study on dollar cost averaging

Abstract: In this presentation I investigate the significance of dollar cost averaging (DCA) strategies in a retirement savings setting. DCA is the practice of systematically investing a fixed amount at regular intervals, with the goal of lowering the average cost per share over time and mitigating the impact of market fluctuations. The study utilizes a rigorous mathematical formulation of an investor's wealth, which can be used to derive various closed form expressions of expected value and variance of the investor's terminal wealth. This research highlights the potential benefits and risk-return tradeoJs of DCA compared to the conventional practice of lump-sum investing. It also provides a deeper understanding of how these strategies perform under diJerent market conditions, particularly during periods of high volatility or downturns. Preliminary results indicate that DCA tends to reduce the volatility of terminal wealth but often leads to slightly lower expected returns. This suggests that while DCA may provide the investor with a smoother investment journey, it comes at the cost of potentially lower terminal wealth, emphasizing the need for investors to balance risk tolerance and long-term goals.

Talk 4

Name: Rhoda Dadzie-Dennis

Title: Empirical Analysis of Climate-Smart Investing

Abstract: Climate change represents one of the most pressing challenges of our time, significantly affecting global financial markets. Both institutional and individual investors are increasingly compelled to address the growing risks posed by climate change by incorporating climate-smart strategies into their portfolios. This research explores the

impact of climate change risk on portfolio performance, focusing on the integration of climate change proxies such as Environmental, Social, and Governance (ESG) scores, E-only scores, UNPRI, Disaster Risk Index (DRI), and Impact Factors. By optimizing portfolio weights using these proxies, we provide an analytical framework to enhance resilience to climate risks under various economic scenarios. Our findings indicate that portfolios integrated with certain climate-related factors outperform traditional portfolios in the long run, offering better protection against regulatory changes, physical risks, and transition uncertainties. This research has practical implications for investors looking to safeguard their portfolios against climate-induced market disruptions.