International Conference on Statistics and Data Science
June 23-25, 2025 Vancouver, Canada

Abstract Brochure

Sponsors

The organizers are grateful for support from the following sponsors:

Simon Fraser University

SFU

SIMON FRASER
UNIVERSITY

Pacific Institute of Mathematical Science

Pacific Institute for the Mathematical Sciences

Canadian Statistical Sciences Institute

CANSSI
INCASS

Organizing Committee

The conference program is organized by Jiguo Cao, Haolun Shi, and Liangliang Wang from Simon Fraser University.

Table of Contents

Adam Kashlak, University of Alberta

Invited Session Organized by Adam Kashlak

Jun 23 15:30-17:00 HC1520

Functional Kalman Filtering and Smoothing for Accelerometer Data

Wearable medical devices are becoming more ubiquitous every year as the cost and size of such devices drops. However, these devices can easily capture millions of data points every day. This necessitates the creation of powerful and novel statistical methods for analyzing such data. In this talk, we propose a fully functional data variant of the classic Kalman filter and smoother. To achieve this, we formulate such dynamic linear models in the framework of general Gaussian measures in locally convex spaces. This results in a mathematically sound extension from multivariate (i.e. finite dimensional) time series data to fully functional (i.e. infinite dimensional) time series data. This general setting allows for the formulation of many specific time series models for functional data. We apply these ideas to smoothing noisy accelerometer data collected from foot and leg movements.

Aishwarya Mandyam, Stanford University

Data-Driven Decision Making in Public Health

Jun 23 10:00-11:30 HC1315

Adaptive Interventions with User-Defined Goals for Health Behavior Change

Promoting healthy lifestyle behaviors remains a major public health concern, particularly due to their crucial role in preventing chronic conditions such as cancer, heart disease, and type 2 diabetes. Mobile health applications present a promising avenue for low-cost, scalable health behavior change promotion. Researchers are increasingly exploring adaptive algorithms that personalize interventions to each person's unique context. However, in empirical studies, mobile health applications often suffer from small effect sizes and low adherence rates, particularly in comparison to human coaching. Tailoring advice to a person's unique goals, preferences, and life circumstances is a critical component of health coaching that has been underutilized in adaptive algorithms for mobile health interventions. To address this, we introduce a new Thompson sampling algorithm that can accommodate personalized reward functions (i.e., goals, preferences, and constraints), while also leveraging data sharing across individuals to more quickly be able to provide effective recommendations. We prove that our modification incurs only a constant penalty on cumulative regret while preserving the sample complexity benefits of data sharing. We present empirical results on synthetic and semi-synthetic physical activity simulators, where in the latter we conducted an online survey to solicit preference data relating to physical activity, which we use to construct realistic reward models that leverages historical data from another study. Our algorithm achieves substantial performance improvements compared to baselines that do not share data or do not optimize for individualized rewards.

Alex Stringer, University of Waterloo

Innovations in Statistical Theory, Design, and Applications

Jun 24 13:30-15:00 HC1325

Semi-parametric benchmark dose analysis with monotone splines

Benchmark dose analysis is a technique in environmental toxicology used by regulatory agencies to determine allowable doses to hazardous substances. A dose-response curve is fit using potentially large observational sets of data, and a confidence interval for the point on the x-axis where the curve drops below a certain threshold is reported. A new method for benchmark dose analysis with semi-parametric dose-response curves is introduced. The method uses monotone splines, de Boor's algorithm and a reflective Newton's method to provide fast and stable inference for large data sets. Confidence intervals are given based on a pivot approach and novel parametric bootstrap that does not require re-fitting the dose-response model, leading to large improvements in speed and accuracy over previous approaches. The method is applied to the study of prenatal alcohol exposure and child cognition using data from six NIH-funded longitudinal cohort studies. Based on joint work with Tugba Akkaya Hocagil, Richard Cook, Louise Ryan, and Sandra and Joseph Jacobson.

Alexandre Bouchard-Côté, University of British Columbia

Bayesian Computational Methods

Jun 23 13:30-15:00 HC1700 Labatt Hall

How to choose an annealing algorithm

Over the years, several algorithms have been developed to tackle normalization constant estimation. A handful of those have passed the test of time thanks to their capacity to beat the curse of dimensionality in many realistic scenarios: on one hand, Annealed Importance Sampling (AIS) and Sequential Monte Carlo (SMC) methods, and on the other, Parallel Tempering (PT) and Simulated Tempering (ST) algorithms. Indeed many recent developments can be contextualized as members of one of these two families of meta-algorithms.

A priori, these two families of algorithms, AIS/SMC versus PT/ST, appear quite distinct and indeed these communities are largely silos. This leads to an important practical question: for a given problem, which annealing algorithm should be recommended? I will present our work toward tackling this question.

Aline Talhouk, University of British Columbia

Advancing Predictive Analytics for Health Outcomes: Addressing Uncertainty, Customization, High-Dimensionality, and Privacy

Jun 24 13:30-15:00 HC1315

Privacy-Preserving Predictions: Federated Learning for Distributed Healthcare Data

Predictive analytics can help derive insights and make data-driven decisions across various domains, including healthcare. However, training machine learning and AI models requires large datasets, and in the context of healthcare, this data is distributed across institutions, which can pose regulatory challenges for data sharing. In recent years, federated learning (FL) has emerged as a transformative approach to machine learning, enabling decentralized data processing while maintaining data privacy and security. By allowing models to be collaboratively trained on distributed data sources without the need for data centralization, federated learning provides a powerful framework for predictive analytics, particularly in sensitive or regulated environments. My talk explores the implementation of predictive analytics with federated learning, examining how FL enables the development of privacy-preserving predictive models across disparate datasets. I discuss key challenges and implementation barriers in real-world scenarios.

Andrew McCormack, University of Alberta

Invited Session Organized by Adam Kashlak

Jun 23 15:30-17:00 HC1520

The Unbiasedness Threshold

Applications of linear algebra in statistics abound, such as those in linear regression and principal components analysis. Moving beyond linearity, the field of algebraic statistics leverages tools from computational algebra and algebraic geometry to solve statistical problems that involve polynomial functions. In this work I examine statistical hypothesis testing for discrete data from an algebraic perspective, with a focus on questions of the existence of unbiased tests. The sample size needed for the existence of a strictly unbiased test, termed the unbiasedness threshold, is shown to be the minimum degree of a polynomial that separates the null and alternative hypothesis sets. In particular, this result implies that null hypothesis sets must be semialgebraic for there to exist a strictly unbiased test. Explicit sample size requirements for various hypotheses in a multinomial model, such as hypotheses of independence in contingency tables, are given. It is demonstrated that upper bounds for the unbiasedness threshold can be found by computing Gröbner bases, and that such upper bounds are tight when all polynomial power functions can be written as sums of squares.

Ann Smith, University of Huddersfield

Modeling in Natural and Physical Sciences and Engineering

Jun 25 13:30-15:00 HC1315

Data driven Engineering Control Systems

Efficient performance of systems and components has a direct impact on carbon emisions, hence the environment in addition to health and safety, and cost considerations is of important consideration. Ensuring systems are sustainably and ethically maintained is a global expectation thus reliable, robust methods of ascertaining machine health most directly and continuously are imperative. Process monitoring systems continue to grow in both complexity and application, often organically as existing configurations are updated. A wealth of data is amassed, which can result in computational overload if all outputs are directly analysed or when incorporated in highly complex models. However, the mass of signal outputs collected also offers variety of approach. A means of establishing computable prognostic models to accurately reflect reciprocating compressor condition, whilst alleviating computational burdens, is essential.

Archer Yi Yang, McGill University

Statistical Machine Learning Methods

Jun 25 10:00-11:30 HC1325

Multivariate Conformal Selection

Selecting high-quality candidates is crucial in drug discovery and precision medicine. While Conformal Selection (CS) ensures uncertainty quantification, it is limited to univariate responses. We propose Multivariate Conformal Selection (mCS), extending CS to multivariate settings using regional monotonicity and multivariate nonconformity scores for conformal p-values, ensuring finite-sample False Discovery Rate (FDR) control. We introduce two variants: one using distance-based scores and another optimizing scores via differentiable learning. Experiments on simulated and real-world data show mCS enhances selection power while maintaining FDR control, making it a robust tool for multivariate selection.

Bahadir Yüzbaşı, Inonu University, Turkey

High-Dimensional Data Analysis

Jun 25 15:30-17:00 HC1700 Labatt Hall

Groupwise Feature Selection in High Dimensional Data

Groupwise feature selection is a powerful tool for handling high-dimensional datasets, particularly when features naturally form groups or clusters. It helps to improve model accuracy, reduce overfitting, and ensure the interpretability of results. In this talk, we present a new groupwise feature selection technique that deals with correlated or structured data that appears in many fields, including genomics, finance, and image processing. In order to assess its performance, we conducted a Monte Carlo simulation study and microarray data from the Gene Expression Omnibus (GEO) repository. Numerical studies show that our suggested method has better prediction errors and a lower false discovery rate compared to the competitors in literature.

Belal Hossain, St. Paul's Hospital Vancouver, The University of British Columbia

Advancing Predictive Analytics for Health Outcomes: Addressing Uncertainty, Customization, High-Dimensionality, and Privacy

Jun 24 13:30-15:00 HC1315

Boosting Predictions with High-Dimensional Administrative Data: A Tuberculosis Case Study

In conventional clinical prediction models, investigators develop their models based on the available predictors in their development sample. Since they are intended for use at the point of care (e.g., during a clinical encounter), there is an incentive for such models to be parsimonious, having strong predictive power with as few predictors as possible. However, health administrative data often do not contain important clinical predictors such as smoking. In the linked databases, a wide range of healthcare variables are often available that can be used for developing a high-dimensional prediction model (hdPM). Given the high dimensionality and without the use of appropriate penalization (e.g., shrinkage) methods, hdPMs are at high risk of overfitting, leading to overly optimistic predictions. In this talk, I will demonstrate how hdPM can compensate for the lack of clinical predictors while improving prediction accuracy without overfitting the model. I will discuss how hdPM could offer a robust approach for risk stratification in epidemiological research compared to a conventional model that relies only on investigator-specified clinical predictors.

Benjamin Bloem-Reddy, University of British Columbia

Invited Session Organized by Adam Kashlak

Jun 23 15:30-17:00 HC1520

Randomization Tests for Conditional Group Symmetry

Symmetry plays a central role in the sciences, machine learning, and statistics. While statistical tests for the presence of distributional invariance with respect to groups have a long history, tests for conditional symmetry in the form of equivariance or conditional invariance are absent from the literature. This work initiates the study of nonparametric randomization tests for symmetry (invariance or equivariance) of a conditional distribution under the action of a specified locally compact group. We develop a general framework for randomization tests with finite-sample Type I error control and, using kernel methods, implement tests with finite-sample power lower bounds. We also describe and implement approximate versions of the tests, which are asymptotically consistent. We study their properties empirically on synthetic examples, and on applications to testing for symmetry in problems from high-energy particle physics.

Bingfan Liu, Simon Fraser University

Statistical Machine Learning Methods

Jun 25 10:00-11:30 HC1325

LongSurvFormer: Transformer-based Joint Modeling for Dynamic Survival Prediction using Longitudinal Images

Survival analysis leveraging longitudinal medical images plays a pivotal role in healthcare, espe- cially for the early detection and prognosis of dis- eases by providing insights beyond single-image assessments. However, current methodologies of- ten inadequately utilize censored data, overlook correlations among longitudinal images measured over multiple times, and lack interpretability. We introduce LongSurvFormer, a novel Transformer- based neural network that integrates longitudi- nal medical imaging with structured data for sur- vival prediction. Our architecture comprises three key components: a Vision Encoder for extract- ing spatial features, a Sequence Encoder for ag- gregating temporal information, and a Survival Encoder based on the Cox proportional hazards model. This framework effectively incorporates censored data, addresses scalability issues, and enhances interpretability through occlusion sen- sitivity analysis and dynamic survival prediction. Extensive simulations and a real-world applica- tion in Alzheimer's disease analysis demonstrate that LongSurvFormer achieves superior predictive performance and successfully identifies disease- related imaging biomarkers.

Boyi Hu, Columbia University

High-Dimensional Data Analysis

Jun 25 15:30-17:00 HC1700 Labatt Hall

TPClust: Temporal Profile-Guided Disease Subtyping Using High-Dimensional Omics Data

One of the primary challenges in treating and preventing neurodegenerative diseases is the extensive heterogeneity in the clinico-pathological state of older individuals and at the molecular level, suggesting the presence of subgroups of individuals who share certain biological features but respond differently to disease risk factors, which makes this essentially a disease subtyping problem. Using the omics data obtained from high-throughput sequencing technologies to perform cluster analysis has been proven to be an effective tool for subtyping diseases and uncovering the complex biological mechanisms behind these subtypes. Traditionally, the disease subtyping analyses often focus on identifying similar patterns in the omics, such as gene expressions. However, the gene expression data is typically high-dimensional and contains many gene sets, some of which may not be associated with the disease, while others may be strongly associated with confounders that are irrelevant to the target disease. The confounding subgroups within these informative gene sets can dominate the clustering process, leading to subgroups that fail to capture clinically meaningful disease subtypes. To identify the subgroups with clear clinical related interpretation, multiple outcome-guided disease subtyping approaches have been developed and shown promise in various medical applications. However, these methods generally assume time-invariant relationships between clinical variables, limiting their abilities to accurately incorporate longitudinal trajectory information and causing a lack of flexibility for gene and pathway selection. This motivates the development of a novel outcome-guided disease subtyping approach tailored for longitudinal studies. In this project, we propose a novel semi-parametric latent mixture regression model to integrate longitudinal clinical data and high-dimensional omics data, denoted as ``TPClust'' method. As a joint latent class model, TPClust can simultaneously identify clinically and biologically meaningful AD subtypes, estimate the time-varying relationships between the AD-related outcome and AD risk factors, and perform feature selection to identify the signal genes and pathways. Numerical studies and a real-world application on Alzheimer's disease subtyping among elderly non-Hispanic Whites demonstrate the superiority of the proposed approach. This work is crucial for advancing personalized medicine, paving the way for more targeted therapeutics, and the development of individualized clinical trials aimed at mitigating AD risk and progression.

Clara Xing Wang, Illinois State University

Novel statistical methods for complex data analysis

Jun 25 10:00-11:30 HC1315

Multi-output Extreme Spatial Model for Complex Production Systems

Most machine learning models model the mean response and are inappropriate for studying abnormal extreme events that are often of the main interest. Engineering applications of extreme models usually focus on individual extreme events, which is insufficient for complex systems with correlations. Moreover, existing extreme spatial models in other domains cannot be directly applied to controllable production systems. In this project, first, we propose an extreme spatial model that facilitates efficient modeling of multi-output response control systems with a bilinear function on two spatial domains for control variables and measurement locations. Marginal parameter modeling and extremal dependence are investigated. In addition, an efficient graph-assisted composite likelihood estimation and corresponding computational algorithms are developed to cope with high dimensional outputs. Its application to composite aircraft assembly shows that the proposed model enables comprehensive analyses with superior predictive performance on extreme events to canonical methods. Also, we explore the improvement of the proposed extreme spatial model and its application in climate risk.

Cédric Beaulac, Université du Québec à Montréal

Statistical Imaging and Vision

Jun 23 13:30-15:00 HC1520

From Pixels to Shapes: A Functional Framework for Image Analysis

In this presentation, we introduce a novel perspective on image analysis that emphasizes objects and their shapes rather than individual pixels. By moving away from pixel-based approaches and analyzing images as collections of objects characterized by both color and contour, we establish a new framework that is more interpretable, less sensitive to resolution changes, and better aligned with the real-world geometry of the objects in images. To achieve this, we propose representing contours using coordinate functions—bivariate functional observations. Representing these functions through Fourier expansion provides an elegant solution to the alignment problem and allows us to apply a wide range of functional approaches to shape analysis, such as multivariate functional principal component analysis. We also discuss how to extend our approach to more complex images containing multiple objects. Finally, we illustrate the proposed method through a range of statistical applications, including sampling (generation), classification, and clustering on real image datasets.

Cristian Oliva Aviles, Genentech Inc.

Contributed Session 4: Wednesday

Jun 25 15:40-17:20 HC1325

Tolerance Intervals for Unbalanced Linear Mixed Models

In various scientific and industrial contexts, there is often a need to compute tolerance intervals for linear mixed models (LMMs) with unbalanced data. Despite this need, the development of tolerance intervals for unbalanced LMMs has predominantly focused on one-way random-effects models. We present an innovative approach for calculating (β, γ)-tolerance intervals for a wider class of unbalanced LMMs, using the concept of Generalized Pivotal Quantities and Monte Carlo sampling techniques. Simulation studies confirm that the tolerance intervals maintain coverage probabilities close to their nominal levels. Further, we showcase the application of our method by estimating the shelf life of biological products using stability data.

Danny Santano, The University of British Columba, Okanagan

Contributed Session 4: Wednesday

Jun 25 15:30-17:00 HC1325

Detecting Financial Market Crashes with Density Ratio Estimation: A Sliding Window Approach for Distribution Shift Analysis

Market crashes are unstable financial periods characterized by substantial shifts in the distribution of asset returns. The detection of these shifts can provide economic insights or early warning signs that could potentially mitigate losses. In this talk, we utilize density ratio estimation (DRE) to identify distributional changes surrounding market crash events. Specifically, we use a relative unconstrained least-squares importance fitting (RuLSIF) DRE method in a double sliding window algorithm that compares return distributions of pre- and post-crash windows, as well as adjacent non-overlapping sliding windows with variable timeframes to capture temporal dynamics at different scales. When applying these methods to cross-sectional equity returns data from the 2008 financial crash, we observe differences in density ratios between unstable and stable market periods. The comparative analysis demonstrates a contrast between stable and crash periods, with elevated density ratios during market disruption that are not present during stable conditions. Our findings demonstrate the efficacy and potential use of DRE approaches as a method for evaluating the stability of a market, particularly as an early warning mechanism for a market crash.

David Stenning, Simon Fraser University

Modeling in Natural and Physical Sciences and Engineering

Jun 25 13:30-15:00 HC1315

Multistage Astrostatistical Analyses

Many astrostatistics data analyses proceed according to a multistage process, whereby the output of a primary analysis is used as the input to a secondary analysis. In such settings, quantifying and carrying forward uncertainty when making inferences and/or predictions is often difficult. In this talk, I will present examples of how multistage analyses can be utilized, often in combination with novel Bayesian modeling and computation, to tackle recent data-analytic challenges in solar and stellar physics.

Dengdeng Yu, University of Texas at San Antonio

Novel statistical methods for complex data analysis

Jun 25 10:00-11:30 HC1315

Word Embeddings via Causal Inference: Gender Bias Reducing and Semantic Information Preserving

With widening deployments of natural language processing (NLP) in daily life, inherited social biases from NLP models have become more severe and problematic. Previous studies have shown that word embeddings trained on human-generated corpora have strong gender biases that can produce discriminative results in downstream tasks. Previous debiasing methods focus mainly on modeling bias and only implicitly consider semantic information while completely overlooking the complex underlying causal structure among bias and semantic components. To address these issues, we propose a novel methodology that leverages a causal inference framework to effectively remove gender bias. The proposed method allows us to construct and analyze the complex causal mechanisms facilitating gender information flow while retaining oracle semantic information within word embeddings. Our comprehensive experiments show that the proposed method achieves state-of-the-art results in gender-debiasing tasks. In addition, our methods yield better performance in word similarity evaluation and various extrinsic downstream NLP tasks.

Deniz Sezer, University of Calgary

Environmental Modelling

Jun 23 10:00-11:30 HC1520

A Markov Chain Gaussian Process framework for modeling wind speed over a large geographical area

An important design consideration when modeling the joint statistical behavior of wind speed measurements over multiple sites and time points is that the models should allow inference for future sites for which no historical measurements are available. Also, the models should be able to combine information from different databases of atmospheric measurements. In this talk, I will introduce a Markov Chain Gaussian process framework which achieves both design goals, describe the estimation procedure for the models considered and give an application to the short term forecasting of wind speed at over 100 weather stations in Alberta.

Divya Shanmugam, Cornell Tech

Data-Driven Decision Making in Public Health

Jun 23 10:00-11:30 HC1315

When Coverage Drives Care: The Public Health Consequences of Insurance Design

The evaluation of clinical care is fundamentally constrained by the need for large, labeled datasets—resources that range from expensive to impossible to obtain. This talk introduces methods to evaluate the quality of clinical care by making use of unlabeled data to evaluate decision-making in two settings: looking backward at human clinical judgment, and forward toward algorithmic systems. In the first part, I focus on retrospective evaluation of past diagnostic decisions, and present a framework to measure the rate of hidden diagnoses in the health record. In the second part, I turn to prospective evaluation of algorithmic decision-making in clinical care, and introduce a method to assess the performance of clinical predictive models in the absence of abundant labeled data. Together, these projects highlight a shared challenge in both human and algorithmic decision-making – evaluating decision quality under data constraints – and provide steps towards robust evaluation of both human and algorithmic care.

Donghui Son, Simon Fraser University

Statistical Imaging and Vision

Jun 23 13:30-15:00 HC1520

Multitask Spatial Bayesian Additive Regression Tree Model with Response-Specific Variable Selection: Application to Imaging Genetics.

This paper introduces Bayesian additive regression trees (BART) for spatially correlated multivariate responses. Existing imaging genetic studies of the Alzheimer's Disease Neuroimaging Initiative (ADNI) mainly focus on the linear association between images of volumetric and cortical thickness values measured by magnetic resonance imaging (MRI) and summarize the structure of the human brain. However, current studies overlook the potential interactions of single nucleotide polymorphisms (SNPs) and the nonlinear relationship between phenotypes and SNPs. We provide a multivariate BART model with a spatial process to address this problem. Furthermore, using a bivariate spatial correlation, our model simultaneously allows the correlation observed in brain neighbor structure within the same hemisphere and the correlation between the left and right hemispheres. The methodology also includes a response-specific variable selection technique to identify important SNPs. Our new BART model demonstrates enhanced prediction performance and easy uncertainty quantification. We illustrate the benefits of our multivariate spatial BART proposal over existing models via simulation studies and application to the ADNI dataset.

Edward Yalley, Louisiana State University

Contributed Session: Monday

Jun 23 15:30-17:00 HC1315

Misconceptions, understandings, and developmental theories in stochastic reasoning

In recent times, the ability to think and reason stochastically forms a vital part of student learning and society. Shaughnessy and Bergman (1993) used the term "stochastic" to imply both probability and statistics (p. 178). Downing (2009) explained statistics as a way of analyzing data. Thus, in statistics, we master the practice of collecting, analyzing, interpreting, and presenting numerical data in massive quantities. Furthermore, one purpose of statistical practice is for inference so that every manipulation contributes "to signify measures of attributes of a situation about which a decision was to be made" (Cobb 1999, p. 13). For instance, emerging fields like data science and educational statistics require a firm foundation in statistics for practitioners. Therefore, educationists like Armah, Aseidu-Addo, and Owusu-Ansah (2016) agree that a course in statistics can enhance students' understanding and use of data when the need arises. As a result, in the USA and other countries like Ghana, a course for (graduate school programs) and a unit (for grade level students) of statistics is needed towards the completion of respective educational levels. However, the works of Hahs-Vaughn and Lomax (2020) have raised some concerns. They revealed that some of the causes of poor grades of students in the field include "not having a quantitative course for some time, apprehension built up by delaying taking statistics, a poor past instructor or course, or less than adequate past success" (p. 2). Their report seems to suggest that having a quantitative course for some time, not delaying taking statistics courses, a helpful instructor or course, and adequate past success can positively influence students' grades in statistics. While scholars like the aforementioned have contributed to the literature of the field, there are existing underlying cognitive obstacles, conceptual structures, and misconceptions that have contributed to the problem (Shaughnessy & Bergman, 1993). There is therefore the need for developmental pathways which teachers can reference and employ for practical classroom teaching and learning. In this study, I present a brief review of selected literature on some misconceptions, and understandings in the teaching and learning of probability and statistics. Moreover, a discussion on some developmental theories and pathways as applied in my own clinical interviews with learners is presented to support my arguments. Thus, at the core of this write-up is a presentation of possible or expected student responses, and a sense of how to leverage student-teacher interactions to promote conceptual development in probability and statistics.

Ehsan Karim, University of British Columbia

Pushing Causal Inference Forward: Blending Machine Learning and Statistical Innovation

Jun 24 10:00-11:30 HC1325

Comparing TMLE Variants: Balancing Efficiency and Reliability in Causal Inference

Background: Targeted Maximum Likelihood Estimation (TMLE) is widely used for causal effect estimation, offering double robustness and asymptotic efficiency. Recent advances—SCTMLE, DCTMLE, CVqTMLE, and full CVTMLE—enhance applicability to complex machine learning models.
Objective: We compare these TMLE variants, with and without repeated sample splitting, against Vanilla TMLE in estimating average treatment effects.
Methods: Using simulations and real-world data, we assessed bias, MSE, coverage, bias-eliminated coverage, and standard error accuracy.
Results: DCTMLE had the lowest MSE; SCTMLE showed lowest bias but higher relative error. Full CVTMLE had best coverage and uncertainty quantification. CVqTMLE balanced performance and efficiency. Repetition improved results, stabilizing after ~25–30 replicates.
Conclusion: DCTMLE is ideal for minimizing error; full CVTMLE for uncertainty. CVqTMLE is practical for constrained settings. Method choice should reflect study goals.

Elizabeth Chou, National Chengchi University, Taiwan

Statistical Learning and Dimensionality Reduction

Jun 23 15:30-17:00 HC1700 Labatt Hall

Improving Neural Network Performance with PCA-Based Dimensionality Reduction

Neural networks have shown remarkable success in modeling complex, high-dimensional data, but their efficiency and interpretability remain ongoing challenges, especially in small-sample and resource-constrained settings. This talk introduces a practical approach that integrates Principal Component Analysis (PCA) into neural network workflows to reduce input or hidden-layer dimensionality while preserving key structural information. The method is particularly effective when applied to similarity-based architectures, such as Siamese Neural Networks, where paired comparisons require compact and discriminative representations. By projecting data into lower-dimensional subspaces guided by explained variance thresholds, we simplify the model structure, reduce overfitting risk, and enhance computational efficiency. This strategy offers a data-driven, model-agnostic enhancement that is especially relevant for applied statisticians and data scientists working with high-dimensional, noisy, or imbalanced datasets.

Erin Zhang, Simon Fraser University

Functional and Longitudinal Modeling

Jun 23 15:30-17:00 HC1325

Robust Bayesian functional principal component analysis

We develop a robustBayesian functional principal component analysis (RB-FPCA) method that utilizes the skew elliptical class of distributions to model functional data, which are observed over a continuous domain. This approach effectively captures the primary sources of variation among curves, even in the presence of outliers, and provides amore robust and accurate estimation of the covariance function and principal components. The proposed method can also handle sparse functional data, where only a few observations per curve are available.We employ annealed sequential Monte Carlo for posterior inference, which offers several advantages over conventional Markov chain Monte Carlo algorithms. To evaluate the performance of our proposed model, we conduct simulation studies, comparing it with well-known frequentist and conventional Bayesian methods. The results show that our method outperforms existing approaches in the presence of outliers and performs competitively in outlier-free datasets. Finally, we demonstrate the effectiveness of our method by applying it to environmental and biological data to identify outlying functional observations.

Fatemeh Mahmoudi, Mount Royal University

New developments on survival analysis and variable selection

Jun 24 10:00-11:30 HC1315

Recent Advances in Variable Selection

Selecting relevant variables is a crucial aspect of building statistical models, as it impacts both their interpretability and predictive accuracy. This paper provides a detailed review of recent advancements in variable selection techniques, examining their applications across different data structures and modeling frameworks. We discuss classical approaches such as stepwise regression and LASSO, alongside newer methodologies, including machine learning-driven strategies and Bayesian variable selection. By assessing these methods in various scenarios, we highlight their strengths, limitations, and trade-offs, offering practical guidance for their implementation. This review aims to support researchers and practitioners in navigating the evolving landscape of variable selection, ultimately contributing to more reliable and effective statistical models.

Fei Wan, Washington University in St Louis

Causal Inference in Observational Studies

Jun 23 13:30-15:00 HC1325

Propensity Score Matching: should we use it in designing observational studies?

Propensity Score Matching (PSM) stands as a widely embraced method in comparative effectiveness research. PSM crafts matched datasets, mimicking some attributes of randomized designs, from observational data. In a valid PSM design where all baseline confounders are measured and matched, the confounders would be balanced, allowing the treatment status to be considered as if it were randomly assigned. Nevertheless, recent research has unveiled a different facet of PSM, termed "the PSM paradox". As PSM approaches exact matching by progressively pruning matched sets in order of decreasing propensity score distance, it can paradoxically lead to greater covariate imbalance, heightened model dependence, and increased bias, contrary to its intended purpose.

Giseon Heo, University of Alberta

Statistical Genetics, Disease, and Population Modeling

Jun 25 15:30-17:00 HC1315

An Application of Persistent Homology to Hidden Markov Models

Using tools from topological data analysis (TDA) for time series analysis has recently become a popular line of research. The combination of these fields has resulted in new methods for quantifying periodicity and distinguishing behavior in time series data. Persistent homology, a particular method in TDA, studies the evolution of topological features in terms of a single index, and is able to capture higher order features beyond the usual clustering techniques. In this talk, we introduce an application of persistent homology to Hidden Markov Models and Hidden semi-Markov Models.

Guilherme Lopes de Oliveira, McGill University and Federal Center for Technological Education of Minas Gerais (CEFET-MG), Brazil

Contributed Session: Monday

Jun 23 15:30-17:00 HC1315

Addressing Underregistration in Epidemiological Data: A Bayesian Random-Censoring Poisson-Logistic Model

Underreporting of disease cases is a recurring challenge in Epidemiology, introducing bias in statistical modeling. Approaches such as random-censoring Poisson models (RCPM) and Poisson-Logistic models (Pogit) can address underreporting but suffer from identifiability issues, requiring strong and specific prior information that restricts their applicability. We propose a Bayesian approach that combines the RCPM and Pogit strategies, offering greater flexibility and requiring less restrictive prior information. A latent model estimates the occurrence of underreporting (censoring), providing a sufficient condition for the identifiability of the Pogit model. Consequently, our approach enables the simultaneous estimation of disease incidence, censoring probability, and the underreporting rate. The methodology is validated in simulated scenarios and applied to infant mortality data from microregions of Minas Gerais State, Brazil. The results highlight the effectiveness of the method in estimating disease incidence, contributing to more accurate epidemiological surveillance and supporting the development of public policies aligned with the UN Sustainable Development Goals for health.

Grace Yi, University of Western Ontario

Robust Statistical Methods for Complex Data Challenges

Jun 23 10:00-11:30 HC1700 Labatt Hall

Function-on-Scalar Linear Regression with Covariate Measurement Error

Function-on-scalar linear regression is widely used to model the relationship between functional responses and scalar covariates. However, its application is often hindered by measurement error, which is common in many real-world datasets. Directly applying standard function-on-scalar regression to error-contaminated data can lead to biased estimates, a problem that is further complicated by the presence of inactive variables. In this work, we propose a new approach that combines a debiased loss function with a sparsity-inducing penalty to simultaneously estimate functional coefficients and select relevant predictors. We develop an efficient computational algorithm with data-driven tuning parameters and establish the asymptotic properties of the proposed estimator. The method's finite-sample performance is evaluated through numerical studies.

Guilherme Augusto Veloso, Fluminense Federal University, Brazil

Bayesian Models for Complex Data Structures

Jun 24 15:30-17:00 HC1700 Labatt Hall

A Bayesian Space-Time Model for Underreported Data: Application to Tuberculosis in Brazilian States, 2000–2022

Tuberculosis (TB) remains a major global public health issue. In Brazil, approximately 80,000 cases per year have been reported in the last decade and, despite the high incidence, there is evidence that the data are underreported. The failure to detect TB cases sustains transmission, prevents effective treatment and underestimates disease magnitude. This study assesses the quality of TB surveillance data across the 27 federative states of Brazil from 2000 to 2022. We adopted a Bayesian hierarchical space-time model that allows the correction of underreporting while estimating the disease incidence rates. We consider factors linked to TB incidence and conduct a sensitivity analysis to model the probability of case reporting over time in each state. Data quality cluster maps illustrate notification trends and highlight regions where targeted efforts are needed to improve surveillance. Our approach supports monitoring and decision-making to advance the UN Sustainable Development Goals, particularly those related to reducing inequalities and promoting health and well-being.

Guilherme Lopes de Oliveira, McGill University and Federal Center for Technological Education of Minas Gerais (CEFET-MG), Brazil

Contributed Session: Monday

Jun 23 15:30-17:00 HC1315

Addressing Underregistration in Epidemiological Data: A Bayesian Random-Censoring Poisson-Logistic Model

Underreporting of disease cases is a recurring challenge in Epidemiology, introducing bias in statistical modeling. Approaches such as random-censoring Poisson models (RCPM) and Poisson-Logistic models (Pogit) can address underreporting but suffer from identifiability issues, requiring strong and specific prior information that restricts their applicability. We propose a Bayesian approach that combines the RCPM and Pogit strategies, offering greater flexibility and requiring less restrictive prior information. A latent model estimates the occurrence of underreporting (censoring), providing a sufficient condition for the identifiability of the Pogit model. Consequently, our approach enables the simultaneous estimation of disease incidence, censoring probability, and the underreporting rate. The methodology is validated in simulated scenarios and applied to infant mortality data from microregions of Minas Gerais State, Brazil. The results highlight the effectiveness of the method in estimating disease incidence, contributing to more accurate epidemiological surveillance and supporting the development of public policies aligned with the UN Sustainable Development Goals for health.

Haiyi Shi, Simon Fraser University

Modeling in Natural and Physical Sciences and Engineering

Jun 25 13:30-15:00 HC1315

The condition numbers of stochastic inverse problems

Data is often collected through measurements or physical experiments conducted at specific locations or times. The selection of when and where plays an important role in accurately inferring parameters of models from the data. For example, how long should we wait before starting the measurement? How frequently should we take measurements— every second, hour, week, or year? In some cases, physical experiments are difficult and costly to perform, which makes understanding the optimal timing and location for data collection even more important for accurate statistical inference. In this work, we derive a new type of condition number under the framework of stochastic inverse problems and demonstrate that it is a useful and accessible tool which provides such crucial information in the data collection process. In theory, it measures the sensitivity of the solution with respect to the error in estimating the distribution of data. The corresponding numerical algorithm is developed and we compare the new one with the classic condition number. Last, we illustrate the utilities of the two condition numbers in two physics-based models.

Hanna Frank, University of British Columbia

Advancing Predictive Analytics for Health Outcomes: Addressing Uncertainty, Customization, High-Dimensionality, and Privacy

Jun 24 13:30-15:00 HC1315

Customizing Risk: Building a Disease-Specific Comorbidity Index for Multiple Sclerosis

Though the utility of comorbidity summary scores is widely recognized, commonly used measures may be too general for many disease contexts. Using a British Columbia (BC) MS cohort with health administrative data (n = 9,856), we are developing predictive models for time to treatment initiation and mortality, which will form the basis of an MS-specific comorbidity summary index (MSCSI). The final index will be externally validated in two independent cohorts from Manitoba and Sweden. The performance of the newly developed index will be compared to that of conventional comorbidity indices (e.g., the Charlson Index) in the MS setting. My talk will highlight the importance of disease- and outcome-specific comorbidity indices and walk you through the various stages of this ongoing project—from selecting comorbidities to include, to presenting initial internal and external validation results.

Hedayat Fathi, Université Laval

Contributed Session 3: Wednesday

Jun 25 13:30-15:00 HC1325

Selection of functional predictors and smooth coefficient estimation for scalar-on-function regression models

In the framework of scalar-on-function regression models -- in which several functional variables are employed to predict a scalar response -- we propose a methodology for selecting relevant functional predictors while simultaneously providing accurate smooth (or, more generally, regular) estimates of the functional coefficients.
We suppose that the functional predictors belong to a real separable Hilbert space, while the functional coefficients belong to a specific subspace of this Hilbert space. Such a subspace can be a Reproducing Kernel Hilbert Space (RKHS). This ensures the desired regularity characteristics, such as smoothness or periodicity, for the coefficient estimates.
Our procedure, called SOFIA (Scalar-On-Function Integrated Adaptive Lasso), is based on an adaptive penalized least squares algorithm that leverages functional subgradients to efficiently solve the minimization problem.
We demonstrate that the proposed method satisfies the functional oracle property, even when the number of predictors exceeds the sample size.
SOFIA's effectiveness in variable selection and coefficient estimation is evaluated through extensive simulation studies and a real-data application to GDP growth prediction.

Henan Xu, University of Waterloo

Statistical Machine Learning Methods

Jun 25 10:00-11:30 HC1325

Functional Causal Mediation Analysis with Sparse Longitudinal Data and a Zero-inflated Count Outcome

Causal mediation analysis plays a crucial role in understanding direct and indirect causal path ways between a treatment and an outcome. Existing methodologies concerning a zero-inflated count outcome primarily rely on simulation-based approaches which are computationally demanding. Fur thermore, established techniques for incorporating a time-varying mediator fail when dealing with sparse and irregularly observed longitudinal data, commonly encountered in health-related studies. In this work, we propose a novel approach for causal mediation analysis with a sparse longitudi nal mediator and a zero-inflated count outcome. Our framework leverages principal component analysis through conditional expectation (PACE) to recover the mediator process and introduces a marginalised functional zero-inflated Poisson model to derive closed-form causal effects. This approach addresses both the computational challenges of simulation-based methods and the limi tations of standard functional mediation techniques in sparse settings. We demonstrate the utility of our method through simulation studies and a real-world application, analysing gender effects on rehospitalisations after coronary artery bypass grafting from a MIMIC-IV dataset.

Hongbin Zhang, University of Kentucky

Emerging Methods in Survival and Longitudinal Data Analysis

Jun 23 10:00-11:30 HC1325

Inferring the timing of antiretroviral therapy by zero-inflated random change point models using longitudinal data subject to left-censoring.

We propose a new random change point model that utilizes routinely recorded individual-level HIV viral load data to estimate the timing of antiretroviral therapy (ART) initiation in people living with HIV. The change point distribution is assumed to follow a zero-inflated exponential distribution for the longitudinal data which is also subject to left-censoring, and the underlying data-generating mechanism is a nonlinear mixed effects model. We extend the Stochastic EM (StEM) algorithm by combining a Gibbs sampler with Metropolis-Hastings sampling. We apply the method to real HIV data to infer the timing of ART initiation since diagnosis. Additionally, we conduct simulation studies to assess the performance of our proposed method.

Hua Liu, Xi'an Jiaotong University

Advanced Statistical Modeling for Complex Time-to-Event and Spatial-Temperal Data

Jun 24 13:30-15:00 HC1700 Labatt Hall

Similarity-Informed Transfer Learning for Multivariate Functional Censored Quantile Regression

To address the challenge of utilizing patient data from other organ transplant centers (source cohorts) to improve survival time estimation and inference for a target center (target cohort) with limited samples and strict data-sharing privacy constraints, we propose the Similarity-Informed Transfer Learning (SITL) method. This approach estimates multivariate functional censored quantile regression by flexibly leveraging information from each source cohort based on its similarity to the target cohort. Furthermore, the method is adaptable to continuously updated real-time data. We establish the asymptotic properties of the estimators obtained using the SITL method, demonstrating improved convergence rates. Additionally, we develop an enhanced approach that combines the SITL method with a resampling technique to construct more accurate confidence intervals for functional coefficients, backed by theoretical guarantees. Extensive simulation studies and an application to kidney transplant data illustrate the significant advantages of the SITL method. Compared to methods that rely solely on the target cohort or indiscriminately pool data across source and target cohorts, the SITL method substantially improves both estimation and inference performance.

Irene Vrbik, University of British Columbia, Okanagan

Innovations in Statistical Theory, Design, and Applications

Jun 24 13:30-15:00 HC1325

Quantitative Insights into Data Science Curricula

The evolving and interdisciplinary nature of Data Science presents unique challenges for curriculum design. Unlike disciplines with long-standing traditions and broadly agreed-upon core content, Data Science exhibits significant variability across institutions and lacks consensus on essential curricular components. Furthermore, as technologies and methodologies continue to evolve, Data Science curricula require ongoing reassessment and adaptation.

This research introduces a multi-phase approach to analyzing and improving Data Science curricula. In the first phase, we apply a topic modelling technique—Latent Dirichlet Allocation (LDA)—to uncover thematic structures in a corpus of course descriptions from Data Science programs across North America. In the second phase, we present a structural framework for quantitatively evaluating and visualizing curricula using directed acyclic graphs (DAGs), which we have developed into an R package called CurricularAnalytics. This tool aids in identifying critical pathways that can be targeted for strategic improvement and pedagogical innovation. In the third phase, we apply LDA to a Data Science job listings corpus to assess the alignment between curricular content and industry demands.

While these techniques have been demonstrated using our own Data Science programs, they offer a generalizable framework for curriculum evaluation and improvement—particularly valuable in disciplines characterized by rapid evolution and conceptual ambiguity.

Jabed Tomal, Thompson Rivers University

Bayesian Models for Complex Data Structures

Jun 24 15:30-17:00 HC1700 Labatt Hall

A Bayesian hierarchical generalized weighted Poisson regression model for analyzing over- and under-dispersed count data: A case study of fertility patterns in Bangladesh

Traditional Poisson regression models are unable to account for over- or under-dispersion in count data, which may lead to incorrect inferences. To address this limitation, we propose a Bayesian Hierarchical Generalized Poisson (BHGP) regression model to analyze data from the Bangladesh Multiple Indicator Cluster Survey (MICS), collected using multistage sampling. In addition to modeling dispersion, the proposed method probabilistically incorporates survey weights, which ensures that inferences are representative at both the population and stratum levels. The hierarchical structure of the model includes both fixed and random effects, which enables the investigation of individual, household, and regional factors that influence fertility. We evaluated model performance using the Bayesian Information Criterion (BIC) and found that the BHGP model outperforms existing Bayesian Poisson regression models. The results are found to be robust across a range of prior distributions, including Normal, Laplace, Cauchy, and Spike-and-Slab, with varying levels of shrinkage. The estimated expected number of children ever born (CEB) in Bangladesh is found to be 2.30, with a 95% confidence interval of (2.283, 2.315). The fertility rate is positively associated with the age of women and their husbands, while higher education levels of both partners, greater household wealth, higher women's age at first marriage, and access to the media are negatively associated with fertility. Furthermore, women in female-headed households and those whose husbands have multiple wives are estimated to have fewer children than their counterparts. From an application perspective, our study contributes to the development of evidence-based policies aimed at improving maternal and child health, promoting family planning, and improving population well-being in Bangladesh.

Jean-François Bégin, Simon Fraser University

Time Series and Financial Modeling

Jun 24 15:30-17:00 HC1325

The stochastic behaviour of electricity prices under scrutiny: Evidence from spot and futures markets

This research proposes a stochastic volatility jump-diffusion model for pricing electricity derivative contracts. The main objective is to develop a model that effectively captures the characteristics and stylized facts of the electricity spot market, such as mean reversion, changing expectations in the spot price's long-run level, seasonality, extreme volatility, price spikes, and time-varying jump intensity. We employ a particle filter that relies on both spot prices and futures data to estimate model parameters. The results demonstrate that incorporating the aforementioned features is crucial for accurately fitting both spot and futures prices, as evidenced by data from the Australian electricity market.

Jeffrey Andrews, University of British Columbia, Okanagan

Modeling in Natural and Physical Sciences and Engineering

Jun 25 13:30-15:00 HC1315

A Binned and Truncated Mixture Modelling Approach for Raman Spectroscopy

A finite mixture model-based approach is introduced for the analysis of raw data arising from Raman spectroscopy experiments which envisions the observed data as relative counts in 'bins'; akin to a complex histogram. We discuss the implementation of an expectation-maximization algorithm for such an approach, along with other important problems like initialization and model selection. We will compare the proposed method with some standard approaches for Raman data, and dive into some nuances in modelling this type of raw data

Jesse Ghashti, University of British Columbia, Okanagan

Contributed Session: Tuesday

Jun 24 15:30-17:00 HC1315

A Kernelized Similarity Learning Framework for Clustering Mixed-Type Data with Applications to Spectral Clustering

Spectral clustering constructs an affinity matrix using a similarity function, then partitions data into k-groups based on the graph Laplacian's eigenstructure. However, existing similarity functions for mixed-type data—data consisting of continuous, nominal, and ordinal variables—can oversimplify proximity, leading to unintuitive partitions that fail to capture underlying clustering structures. We propose a kernelized similarity learning framework that quantifies proximity within each variable type and weights their contributions to the overall similarity calculation based on variable relevancy in capturing clustering structures. Our recent work explores an optimization procedure for kernel bandwidth selection by maximizing the eigengap between the k-th and (k+1)-th eigenvalues of the graph Laplacian. Simulated and real data analyses demonstrate that this framework improves spectral clustering accuracy over existing mixed-type distance and similarity functions.

Jiahua Chen, University of British Columbia

Robust Statistical Methods for Complex Data Challenges

Jun 23 10:00-11:30 HC1700 Labatt Hall

Byzantine--tolerant distributed learning of finite mixture models

Traditional statistical methods must evolve to keep pace with modern distributed data storage systems. One widely used strategy is the split-and-conquer framework, where models are trained independently on local machines and their parameter estimates are then averaged. While effective for many problems, this approach breaks down when applied to finite mixture models due to the label switching problem---the arbitrary permutation of subpopulation labels across local machines. To tackle this, Mixture Reduction (MR) methods have been proposed, which help reconcile label mismatches. However, MR techniques are not resilient to Byzantine failures, where some local machines may send highly erroneous or even malicious outputs. This paper presents Distance Filtered Mixture Reduction (DFMR), a robust and efficient extension of MR designed to withstand Byzantine failure. DFMR introduces a novel filtering mechanism based on the distribution of local estimates. By computing pairwise L2 distances between these estimates, DFMR identifies and discards outliers likely caused by corruption, while preserving the integrity of the majority. We offer theoretical guarantees for DFMR, demonstrating that it achieves the optimal convergence rate and is asymptotically equivalent to the global maximum likelihood estimate under standard conditions. Through experiments on both simulated and real-world datasets, we show that DFMR delivers reliable and accurate results even in adversarial environments.

Jianghu(James) Dong, University of Nebraska Medical Center

Causal Inference in Observational Studies

Jun 23 13:30-15:00 HC1325

Dynamic Biomarker Regime Switches in Personalized Treatment Strategies

Precision medicine efforts aimed at tailoring therapy to individual patients necessitate flexible statistical methodologies capable of modeling subjects-specific treatment effects and disease trajectories. We propose a method to detect individual-level structural changes in longitudinal modeling to characterize patient-specific biomarker evolution while allowing for partial pooling through hierarchical priors. The joint modeling structure also facilitates integration of time-to-event outcomes. We apply the proposed methodology to a cancer dataset. This work demonstrates the utility of hierarchical change point models for estimating information across subjects while preserving individual variation.

Jiaqi Men, Shanghai University of Finance and Economics, China

Statistical Machine Learning Methods

Jun 25 10:00-11:30 HC1325

Generalized Functional Additive Nonlinear Models with Multimodal Interaction Effects

The effect of temperature on electricity consumption may depend on the wealth of households. Motivated by this electricity consumption inequality issue, we propose a generalized functional additive nonlinear model using functional principal component analysis (FPCA). This model considers the interaction among functional and scalar covariates, which we refer to as multimodal interaction in this article. By combining quasi-likelihood and B-spline approximation techniques, we develop an approach to estimate the proposed generalized Functional Additive Nonlinear model with Multimodal Interaction effects (FANMI) and establish the optimal convergence rate and asymptotic normality of the resultant estimators. We then propose two hypothesis testing procedures to assess the goodness-of-fit of the proposed FANMI model and determine whether some of the bivariate nonparametric functions can be simplified to univariate ones. We derive the asymptotic distributions of the proposed test statistics. Multiple simulation studies are conducted to evaluate the performance of the proposed estimation method and the two hypothesis tests. Finally, the FANMI model and the hypothesis tests are demonstrated by analyzing the interaction effect of temperature and GDP on electricity consumption and the electricity Gini coefficient.

Jiatao Zhong, University of British Columbia, Okanagan

Contributed Session 4: Wednesday

Jun 25 15:30-17:00 HC1325

Novel statistical methods for complex data analysis

This talk proposes an energy-based segmentation method, facilitated by the change point detection. We apply the Kullback-Leibler (KL) divergence to demonstrate the feasibility of our method for non-Gaussian noised images.

Jiatao Zhong and Xiaoping Shi (co-speakers), University of British Columbia-Okanagan

Novel statistical methods for complex data analysis

Jun 25 10:00-11:30 HC1315

Energy-based segmentation methods for non-Gaussian noised images

Abstract: We propose an energy-based segmentation method, facilitated by the change point detection. We apply the Kullback-Leibler (KL) divergence to demonstrate the feasibility of our method for non-Gaussian noised images. Notably, the algorithm for this model automatically determines whether the model is solvable by Gaussian approach and seamlessly transits from a Gaussian solution to a non-Gaussian alternative, and the method can also automatically determine the optimal number of classifications. Furthermore, due to its iterative nature, it can detect and segment small regions within an image that are undetectable by other methods. In comparison to the traditional maximum between-class variance method, for bimodal grayscale images, this method provides improved thresholding accuracy. Additionally, in the context of multiple threshold identification, the proposed method outperforms techniques such as K-Means++, Gaussian mixture models, and Adaptive Thresholding when segmenting multimodal grayscale images of breast cancer, cell, and wildfire images.

Jie Jian, University of Chicago

Bayesian Models for Complex Data Structures

Jun 24 15:30-17:00 HC1700 Labatt Hall

Bayesian Tensor Decomposition for Uncovering Complex Dependencies in International Trade

We present a Bayesian tensor factorization model for inferring latent group structures from dynamic pairwise interaction patterns. International trade has evolved over the past decades under globalization, while recent years have seen shifts that complicate trade dynamics. Trade data are complex, varying over time, across different commodities, and among multiple countries, which makes them difficult to analyze. They are often sparse, with large variations in nonzero values. To address these challenges, we represent international trade data as a tensor and develop a Bayesian Poisson tensor factorization model to extract a low-dimensional, interpretable representation of underlying trade patterns.

Jingxue Feng, Simon Fraser University

Statistical Genetics, Disease, and Population Modeling

Jun 25 15:30-17:00 HC1315

A Switching State-Space Transmission Model for Tracking Epidemics and Assessing Interventions

The effective control of infectious diseases relies on accurate assessment of the impact of interventions, which is often hindered by the complex dynamics of the spread of disease. A Beta-Dirichlet switching state-space transmission model is proposed to track underlying dynamics of disease and evaluate the effectiveness of interventions simultaneously. As time evolves, the switching mechanism introduced in the susceptible-exposed-infected-recovered (SEIR) model is able to capture the timing and magnitude of changes in the transmission rate due to the effectiveness of control measures. The implementation of this model is based on a particle Markov Chain Monte Carlo algorithm, which can estimate the time evolution of SEIR states, switching states, and high-dimensional parameters efficiently. The efficacy of the proposed model and estimation procedure are demonstrated through simulation studies. With a real-world application to British Columbia's COVID-19 outbreak, the proposed switching state-space transmission model quantifies the reduction of transmission rate following interventions. The proposed model provides a promising tool to inform public health policies aimed at studying the underlying dynamics and evaluating the effectiveness of interventions during the spread of the disease.

Joan Hu, Simon Fraser University

Advanced Statistical Modeling for Complex Time-to-Event and Spatial-Temperal Data

Jun 24 13:30-15:00 HC1700 Labatt Hall

Learning from Terror Attacks in South Asia with Extended Hawkes Process Models

Terrorism remains a critical global issue, with terrorism-related deaths rising by 22% to 8,352 in 2024, despite fewer overall attacks(https://www.visionofhumanity.org/maps/global-terrorism-index/#/). South Asia, particularly Pakistan, India, and Afghanistan, continues to experience significant terrorist activities. This study leverages the Global Terrorism Database (GTD), which includes records over 192,200 terror attacks from 1970 to 2018. We explore spatio-temporal patterns of the terrorism since 2000 in the three South Asian countries by extended Hawkes process models. The analysis displays the feature of spatio-temporal neighbourhood-stimulating in terror attacks. Future studies include to focus on specific terrorist organizations and explore transnational terrorism by examining cross-border attack dynamics. This is joint work with S.Y. Park and Sunny Wang.

John Braun, University of British Columbia, Okanagan

Statistical Learning and Dimensionality Reduction

Jun 23 15:30-17:00 HC1700 Labatt Hall

Iterated Data Sharpening in Local Linear Regression

Data sharpening in kernel regression has been shown to be an effective method of reducing bias while having minimal effects on variance. Earlier efforts to iterate the data sharpening procedure have been less effective, due to the employment of an inappropriate sharpening transformation. In this presentation, we describe an iterated data sharpening algorithm which does reduce the asymptotic bias at each iteration. Theory and computation show that the new iteration produces interesting boundary effects. The resulting kernel regressions are also less sensitive to bandwidth choice, and data sharpening with data-driven bandwidth selection via cross-validation can lead to more accurate regression function estimation.

This is joint work with Xiaoping Shi, UBCO, and Hanxiao Chen, a PhD student at Brown University.

Julie Zhou, University of Victoria

Innovations in Statistical Theory, Design, and Applications

Jun 24 13:30-15:00 HC1325

Computing and verifying E-optimal regression designs on discrete design spaces

Equivalence theorems are helpful for verifying optimal approximate regression designs. They have been derived for various optimality criteria including A-, c-, D-, E-, and I-optimality criteria. They can be verified easily for any regression model for many optimality criteria, except for E-optimality. In this talk, we show an alternative equivalence theorem for E-optimality, which can be easily verified for any regression model with a discrete design space. We will present detailed numerical methods for finding E-optimal designs and verifying the equivalence theorem, and several examples are given.

Justin Holman, Colorado State University - Pueblo

Contributed Session 3: Wednesday

Jun 25 13:30-15:00 HC1325

Teaching Multiple Regression: A Review

Data analytics is becoming an increasingly common required skill for many undergraduate degree programs. Because data analytics and statistics have many interdisciplinary applications, most undergraduate statistics courses include students with a wide variety of experiences and propensity for the subject matter. Therefore, instructors must consider which pedagogical approaches and learning models have the highest rate of successfully reaching a varied group of students. This is particularly true for the more challenging topics involved in data analytics, such as multiple regression.
Many broad pedagogical approaches to multiple regression have been studied in the context of teaching data analytics. The success of these approaches is typically measured by metrics of student engagement and achievement. The typical structure of this type of study is to identify a pedagogical strategy and describe its execution through class activities and coursework. Then, if possible, the author will analyze student data to identify success, draw conclusions, and suggest further research. Most of these studies provide strategies designed to address a particularly challenging topic or course objective. These typically include the real-world statistical applications of multiple regression.
In order to better understand effective strategies for teaching the concept of multiple regression in this context, existing literature on the topic was analyzed and categorized into broad themes, with a particular emphasis on the case study method and computer applications & tools. The purpose of this paper is to describe recent strategies that have been successfully used to support student outcomes in teaching regression in data analytics courses. This work is descriptive in nature. We do not formally study or compare any of the teaching methods. Instead, this piece is intended to identify the pedagogical methods used to effectively teach multiple regression so that others may apply them to their curricula.

Ke Li, Simon Fraser University

Statistical Learning and Dimensionality Reduction

Jun 23 15:30-17:00 HC1700 Labatt Hall

Rethinking Regression: Insights from Machine Learning

Regression problems arise every time one would like to predict a continuous-valued variable, be it the colour of a pixel, a 3D position, a system configuration or a feature vector. It is well known that regression with square loss yields the conditional mean as the prediction. This is undesirable when there could be many predictions that are all correct, since the conditional mean would effectively average over these predictions and could be far from any of them. As an example, when the prediction takes the form of an image, the conditional mean tends to be blurry and desaturated. On the other hand, in classification problems, ambiguity in labels does not cause an issue because classifiers produce a distribution over class labels as output. Is it possible to get the best of both worlds? In this talk, I will show how to do so using a simple technique, known as conditional Implicit Maximum Likelihood Estimation.

Ken Peng, Simon Fraser University

Environmental Modelling

Jun 23 10:00-11:30 HC1520

Evaluating COVID-19 Hospitalization Risk Using Wastewater Viral Signals: A Multi-State Model Approach

Wastewater surveillance has emerged as a critical tool for tracking the burden of infectious diseases, offering early insights into trends such as COVID-19 hospitalizations. In this presentation, I will discuss recent stochastic process modeling approaches that investigate the association between COVID-19 hospitalizations and wastewater viral signals. These include extensions of distributed lag models with random time lags, as well as joint models linking infections, viral signals, and hospitalizations. A collection of wastewater viral signals and hospitalization records from Ottawa, Canada, is used to illustrate the methods. The proposed approaches are broadly applicable for monitoring other infectious diseases through wastewater surveillance systems.

Kunj Guglani, Australian National University

Contributed Session: Monday

Jun 23 15:30-17:00 HC1315

Comparative Analysis of Privacy in Sampling Methods

Privacy guarantees in data sampling are crucial in preserving sensitive information while ensuring data utility. This study explores the differential privacy (DP) implications of three widely used sampling techniques—simple random sampling (SRS), stratified sampling (SS), and cluster sampling (CS). While these methods serve distinct purposes in statistical analysis, their impact on privacy remains underexplored. We implement SRS, SS, and CS on a dataset and apply DP mechanisms such as Laplace, Gaussian, and Exponential noise to analyze privacy leakage. Key metrics, including privacy budgets (ε, δ) and privacy loss, are empirically evaluated. Additionally, we assess how sampling fraction, cluster sizes, and stratification criteria influence privacy guarantees. Data utility loss is measured through accuracy-based metrics such as mean squared error and classification accuracy to quantify the trade-off between privacy and utility. Our findings reveal that different sampling methods introduce varying degrees of privacy leakage, with stratification potentially improving privacy at the cost of utility. The study highlights the importance of selecting an optimal sampling approach based on privacy and accuracy constraints. These insights contribute to privacy-aware sampling strategies in real-world applications, guiding practitioners in balancing privacy protection with data usability.

Lang Wu, University of British Columbia

Variance Estimation and Statisical Inference

Jun 24 10:00-11:30 HC1700 Labatt Hall

Jointly Modelling Means and Variances in Mixed Effects Models for Efficient and Robust Inference

In longitudinal data analyses, the within-individual repeated measurements often exhibit large variations and these variations appear to change over time. Understanding the nature of the within-individual systematic and random variations allows us to conduct more efficient statistical inferences. Motivated by HIV viral dynamic studies, we considered a nonlinear mixed effects (NLME) model for modeling the longitudinal means, together with a model for the within-individual variances which also allows us to address outliers in the repeated measurements. Statistical inference was then based on a joint model for the mean and variance, implemented by a computationally efficient approximate method. Extensive simulations evaluated the proposed method. We found that the proposed method produces more efficient estimates than the corresponding method without modeling the variances. Moreover, the proposed method provides robust inference against outliers. The proposed method was applied to a recent HIV-related dataset, with interesting new findings. This is joint work with Qian Ye and Viviane D. Lima.

Lawrence McCandless, Simon Fraser University

Pushing Causal Inference Forward: Blending Machine Learning and Statistical Innovation

Jun 24 10:00-11:30 HC1325

Bayesian Quantile Regression for Robust Treatment Effect Estimation

We examine Bayesian quantile regression for estimating quantile treatment effects. The method is applied in a data example that estimates the effect of chronic medical conditions on depression symptoms in Canadian adolescents. This data is well-suited to demonstrating the properties of quantile regression because it has an unusual outcome variable that is measured on an interval scale but taking discrete values. We show that Bayesian quantile regression may give dramatically different results compared to conventional quantile regression. This occurs because the point estimator from conventional quantile regression is calculated using the simplex algorithm, which is affected by discreteness of the data. In contrast, Bayesian quantile regression explores a continuous range of values for the unknown model parameters. The advantage of inference using the full posterior distribution is illustrated using coverage rates for quantile regression predictive intervals. The results demonstrate that inference based on the full posterior distribution for unknown parameters will often yield a better overall fit for the data compared to conventional quantile regression.

Li Xing, University of Saskatchewan

Advanced Statistical Modeling for Complex Time-to-Event and Spatial-Temperal Data

Jun 24 13:30-15:00 HC1700 Labatt Hall

Concurrent Prediction of Multiple Survival Outcomes with a Refined Stacking Algorithm

Xing et al. (2019) developed prediction algorithms, termed multi-task prediction algorithms using revised stacking (MTPS), to enable us conduct the concurrent prediction for multiple outcome variables with high-dimensional predictors integrated into the prediction process. Their algorithms employed the strategy of the stacking algorithm to construct a multi-task learner through a two-step procedure, where separate single learners are constructed in Step 1, and mutually carried information among those learners is then facilitated in Step 2. While their methods have the flexibility in handling both continuous and binary outcomes as well as a mix of them, their methods are not applicable to the context of survival data, which arise commonly in applications yet their analysis is typically challenged by the prevalent issue of censoring.

Expanding their work to handle the prediction of multiple survival outcomes, we develop a new concurrent prediction algorithm by utilizing the revised residual stacking framework, where the parametric accelerated failure time (AFT) model and Elastic Net AFT model are employed. Through simulation studies and a real data application, we demonstrate that the novel enhancement of MTPS for survival outcomes surpasses the performance of their single learners. Consequently, this new refinement MTPS is recommended to researchers in modelling comorbidity diseases. To ensure broader applicability, we now update the original R package, MTPS, to encompass three types of outcomes: continuous, binary, and survival outcomes. This research offers a new dimension to MTPS, allowing a diverse array of applications spanning various domains.

Lin Zhang, Simon Fraser University

Statistical Genetics, Disease, and Population Modeling

Jun 25 15:30-17:00 HC1315

Allele-frequency estimation and ancestry informative marker identification via retrospective regression

Allele frequency estimation at a genetic marker plays a pivotal role in genetic studies. The accuracy of allele frequency estimation impacts the accuracy and power of a genome-wide association study (GWAS). Moreover, allele frequency may differ between seemingly similar populations, which makes allele frequency estimation particularly important for identifying ancestral informative markers (AIMs). Yet, existing allele frequency estimation methods mostly rely on independent sample from a homogeneous population and cannot provide closed form solutions for the maximum likelihood estimator (MLE) of the allele frequencies. To address these challenges, we propose a retrospective regression framework that takes genotype as the response variable, and population and other covariates as the dependent variable. The regression nature of our proposed method enables it to estimate allele frequency in heterogeneous populations and accommodate sample correlation. We support our analytical findings using the 1000 Genome Project genotype data of five super-populations.

Jesus E. Vazquez, University of North Carolina at Chapel Hill

Contributed Session 3: Wednesday
Jun 25 13:30-15:00 HC1325

Robust Estimation for Longitudinal Models Indexed by Time to Event

Shayan Razmi, University of British Columbia

Contributed Session 3: Wednesday
Jun 25 13:30-15:00 HC1325

Prediction of Alzheimer's disease using CNN networks

Liqun Wang, University of Manitoba

Variance Estimation and Statisical Inference

Jun 24 10:00-11:30 HC1700 Labatt Hall

Regularized Estimation of covariance matrix and error variance in high-dimensional models

Estimation of high-dimensional covariance matrix is one of the fundamental and important problems in multivariate analysis and has a wide range of applications in many fields. We present a novel method for sparse covariance matrix estimation via solving a non-convex regularization optimization problem. We establish the asymptotic properties of the proposed estimator and develop a multi-stage convex relaxation method that guarantees any accumulation point of the generated sequence is a first-order stationary point of the non- convex optimization. Moreover, the error bounds of the first two stage estimators of the multi-stage convex relaxation method are derived under some regularity conditions. Numerical results show that our estimator has high degree of sparsity and outperforms the state-of-the-art estimators. We also present a regularized method for estimation of the error variance in high-dimensional linear models. This is joint work with X. Wang and L. Kong.

Lloyd T. Elliott, Simon Fraser University

Clinical Study Design and Meta-Analysis

Jun 25 13:30-15:00 HC1700 Labatt Hall

Statistical considerations for consortia and meta-analysis

HostSeq is a genomic and clinical databank providing whole-genome sequencing and outcomes for 10,000 Canadians. This consortium constitutes Canada's contribution to the global effort to understand the host genetics of COVID-19. I will present HostSeq genetic association results, highlighting the replication of the genes LZTFL1/SLC6A20 and FOXP4 as associated with severe COVID-19. HostSeq is a project-of-projects, with 15 studies contributing. In some cases, the same individual was recuirted by multiple studies, and in one case, a study submitted summary statistics to the Host Genetics Initiative (a global meta-analysis of COVID-19 host genetics studies) independently of HostSeq as a whole. This consortium structure provides a new challenge for meta-analysis in which studies with overlapping subjects are combined. I will describe a new method for correcting bias in meta-analysis due to sample overlap, inspired by the structural challenges in HostSeq. I will describe follow-up work on HostSeq, including serological profiling indicating that vaccination after SARS-CoV-2 infection leads to consistently higher antibody levels, and a genetic association study on long COVID.

Longhai Li, University of Saskatchewan

Bayesian Models for Complex Data Structures

Jun 24 15:30-17:00 HC1700 Labatt Hall

Z-residuals for Diagnosing Bayesian Hurdle Models

Residual diagnostics play a crucial role in frequentist statistics, helping assess model goodness-of-fit, detect model misspecifications, and identify outliers. However, their use in Bayesian statistics remains limited. Recently, Z-residuals have demonstrated superior performance in diagnosing a wide range of frequentist models with non-continuous responses, including censored and count regression models. In this talk, we extend Z-residuals to Bayesian model with a focus on Bayesian hurdle models. Through simulation studies, we show that Z-residual diagnostics are well-calibrated, more powerful, and more informative than widely used posterior predictive checks. Our simulation studies also evaluate different approaches to computing residuals for Bayesian models. Finally, we demonstrate the practical utility of Z-residuals in diagnosing Bayesian hurdle models using real public health datasets.

Longlong Huang, University of the Fraser Valley

New developments on survival analysis and variable selection

Jun 24 10:00-11:30 HC1315

Survival Outcomes Associated with First and Second-Line Palliative Systemic Therapies in Patients with Metastatic Bladder Cancer --- An Application of Restricted Mean Survival Time Analysis

The Cox proportional hazards (PH) model is widely used in medical fields, such as oncology, diabetes, and cardiology, etc. Cox PH regression requires the proportional hazards assumption, that is the hazard ratios between comparing groups is constant over the entire study period. However, this scenario is rarely achieved in the real world. In addition, the hazard ratio simply quantifies the relative difference in risk based on a model-based approach, the absolute magnitude between treatment groups is unclear. Restricted mean survival time (RMST) analysis is considered as a supplement to the Cox PH model. RMST is defined as the area under the survival curve up to a specific time point. It can be understood as the average survival time or life expectancy during a defined time period, which is a straightforward and clinically meaningful way to interpret the contrast in survival between groups. Attracted by these properties, we apply RMST analysis to the metastatic bladder cancer data collected from the BC Cancer Agency and obtain the survival outcomes associated with first and second-line palliative systemic therapies in patients with metastatic bladder cancer.

Louis Arsenault-Mahjoubi, Simon Fraser University

Contributed Session: Tuesday

Jun 24 15:30-17:00 HC1315

Computational methods for deterministic nonlinear non-Gaussian filtering in finance

Financial market dynamics exhibit nonlinearities, dis- continuities or jumps, and are best modelled with several latent variables. Researchers often use particle filtering to estimate the likelihood function of these complicated models as it is flexible, simple to implement, and avoids the curse of dimensionality. However, recent work demonstrates that, in many practical scenarios, deterministic filters based on numerical integration offer faster and smoother estimates of the likelihood function. Integration by parts-based methods are able to accelerate numerical integration in deterministic filters. Yet, integration by parts has only been applied to limited modelling frameworks in the past. In this talk, I will present a generalization of these filtering methods. I then use the generalized filter to accelerate the estimation of models that allow for the leverage effect and multiple latent factors — both key empirical features of market data— in simulation and empirical studies.

Marissa Reitsma, Stanford University

Data-Driven Decision Making in Public Health

Jun 23 10:00-11:30 HC1315

Optimizing public health responses to the syndemic of substance use, overdose, HIV, and hepatitis C virus

Syndemics are interacting clusters of diseases within populations, often driven by social and structural factors. Multi-disease models with realistic behavioral drivers are necessary to quantify the costs and benefits of interventions that address syndemics. We developed a stochastic network-based multi-disease model to evaluate strategies to address the syndemic of substance use disorder, overdose, human immunodeficiency virus (HIV) infection, and hepatitis C virus (HCV) infection among people who inject drugs (PWID). We used separable temporal exponential-family random graph models to simulate interacting sexual and injection equipment-sharing partnerships and capture heterogeneity in transmission risks. After parameterizing and calibrating the model with empirical data to represent an urban PWID network, we simulated the short-term and long-term impacts of scaling single and combined interventions that have three mechanisms of action: 1) reducing the probabilities of infection transmission, 2) increasing drug injection cessation rates, and 3) treating infections. We found that combinations of interventions, implemented to achieve ambitious levels of coverage, were required to meet national goals to decrease HIV and HCV infection incidence by 90% over ten years. Overall, our study underscores the potential value of implementing and expanding syndemic-focused evidence-based comprehensive treatment and harm reduction services for PWID.

Mathias Lécuyer, University of British Columbia

Pushing Causal Inference Forward: Blending Machine Learning and Statistical Innovation

Jun 24 10:00-11:30 HC1325

Training Causal Time-Series Models for Generalizable Forecasting

Deep learning time-series models are often used to make forecasts that inform downstream decisions. Since these decisions can differ from those in the training set, there is an implicit requirement that time-series models will generalize outside of their training distribution. Despite this core requirement, time-series models are typically trained and evaluated on in-distribution predictive tasks. We extend the orthogonal statistical learning framework to train causal time-series models that generalize better when forecasting the effect of actions outside of their training distribution. To evaluate these models, we leverage Regression Discontinuity Designs popular in economics to construct a test set of causal treatment effects.

Matthew Parker, Simon Fraser University

Functional and Longitudinal Modeling

Jun 23 15:30-17:00 HC1325

Bayesian Inference on Time-Varying Parameters in Recapture Models in the Framework of Functional Data Analysis

We have developed a Bayesian inference method for estimating time-varying parameters in conventional recapture models in a new framework of functional data analysis. Our framework uses cubic b-spline basis functions with adaptive smoothing for a fully Bayesian procedure to estimate functional parameters. We apply this framework to the classical Cormack-Jolly-Seber mark-recapture model. We demonstrate that our methods improve the precision of time-varying parameter estimates in a case study on Cinclus cinclus population sizes. Our simulation results indicate that our functional model out-performs the classical model both when the underlying parameters are constant, and when the parameters are varying with time. We further illustrate the applicability of our framework to other ecological models using a batch-marking model as an example.

Maurice O'Connell, University of Manchester

Clinical Study Design and Meta-Analysis

Jun 25 13:30-15:00 HC1700 Labatt Hall

Causal machine learning methods and sequential target trial emulation for dynamic and static treatment strategies deprescribing medications in a polypharmacy population using electronic health records.

Despite recent advances, limited evaluation and guidance are available on the implementation of causal inference in the area of polypharmacy where high-dimensional confounding and medication interactions are present. The DynAIRx project (Artificial Intelligence for dynamic prescribing optimisation and care integration in multimorbidity) aims to develop statistical tools supporting GPs and pharmacists to find patients living with multimorbidity and polypharmacy who might be offered a better combination of medicines. We estimate treatment effects of discontinuing medications in a polypharmacy population. Methods: Within a polypharmacy cohort, we estimate the average causal effect of both time-fixed and time-varying treatment strategies comparing deprescribing versus continuing specific medications as advised by expert clinicians, e.g. individuals stopping either antiplatelets or anticoagulants, neither or both and the corresponding risk of strokes, bleeds and death. We underwent a detailed causal elicitation process with expert clinicians to draw causal diagrams, pre-specify dynamic treatment strategies and identify all important variables from electronic health records (EHRs). We emulate target trials (using both sequential target trials and landmarking approaches) to estimate the average effect of treatment strategies using EHRs from the Clinical Practice Research (CPRD) Database. We target different causal estimands (e.g., total, direct, or separable effects) to allow for competing events from parametric pooled-over time logistic models using G-methods. Negative control outcomes are used to check robustness. Causal machine learning is used to semi-parametrically include a larger combination of medications and interactions than included in our expert elicited causal diagram e.g., Targeted Maximum Likelihood Estimators, augmented inverse probability weighting with data adaptive approaches, cross-fitting, and super learner ensemble learning. We plan to estimate individualised heterogeneous treatment effects (conditional average treatment effects over the smallest subgroups that can be supported by the data), e.g., R-, S-, T-, U-, X-, RS-, DR-learners evaluated with impact fraction rank-weighting metrics. When selecting patients for medication review, we can prioritise patients who are at highest risk of harms (benefits) from (changes in) treatment strategies. Results We aim to present preliminary interim results from a large CPRD database consisting of millions of EHRs. Discussion How do we give better advice in medication reviews to those with multimorbidity and polypharmacy, traditionally excluded from clinical trials? In clinical practice these prescribing decisions have been experienced a large number of times in EHRs. DynAIRx combines causal AI, guidelines and computing power linked to EHRs and where possible randomised controlled trials to estimate these complex causal effects.

Michelle Miranda, University of Victoria

Statistical Imaging and Vision

Jun 23 13:30-15:00 HC1520

Tensor basis strategies for fMRI: fast and scalable MCMC methods

Task-evoked functional magnetic resonance imaging (fMRI) studies play a key role in mapping brain activation patterns, but their analysis within a Bayesian framework presents computational and methodological challenges due to high-dimensional, spatially and temporally correlated data. We propose a Bayesian function-on-scalar model to estimate population-level activation maps for working memory tasks. At the subject level, a composite basis strategy integrates spatial dependencies across voxels and distant regions of interest while accounting for long-memory temporal correlation. At the population level, a canonical polyadic (CP) tensor decomposition is employed to extract shared and subject-specific spatial features, allowing for a structured representation of activation patterns. This hierarchical approach enables computationally scalable posterior inference via MCMC and reveals distinct activation signatures associated with working memory tasks.

Mohsen Sadatsafavi, University of British Columbia

Advancing Predictive Analytics for Health Outcomes: Addressing Uncertainty, Customization, High-Dimensionality, and Privacy

Jun 24 13:30-15:00 HC1315

Quantifying Uncertainty's Clinical Impact: Decision Theory in Predictive Analytics

The contemporary approach to uncertainty quantification in clinical prediction modeling is precision-driven, using inferential statistics targeting pre-specified precision of performance estimates (e.g., 95% CI around c-statistic or calibration slope). However, the relevance of classical inference to clinical decision-making is uncertain. Since prediction models guide patient care, uncertainty can be assessed by its impact on decision outcomes. From a decision-theoretic perspective, uncertainty is linked to loss of value due to potentially suboptimal decisions. This talk reviews recent advances in applying decision theory, particularly Value of Information (VoI) analysis, to risk prediction modeling. We discuss VoI in model development and validation, and key metrics—expected value of perfect information (EVPI) and expected value of sample information (EVSI). We review computational challenges with large administrative or EMR data, and potential solutions based on the central limit theorem and meta-modeling. VoI may also help address fairness in risk stratification. A value-driven approach can complement inferential methods when clinical utility is the focus.

Mohsen Sadatsafavi, The University of British Columbia

Clinical Study Design and Meta-Analysis

Jun 25 13:30-15:00 HC1700 Labatt Hall

Bayesian sample size calculations for external validation studies of risk prediction models

Background: Contemporary sample size calculations for external validation of risk prediction models require users to specify fixed values of assumed model performance metrics alongside target precision levels (e.g., 95% CI widths). However, due to the finite samples of previous studies, our knowledge of model performance in the target population is uncertain, and so choosing fixed values represents an incomplete picture. As well, for net benefit (NB) as a measure of clinical utility, the relevance of conventional inference is doubtful. A Bayesian approach enables proper incorporation of parameter uncertainty and facilitates the use of novel decision-theoretic metrics when planning external validation studies. Methods: We propose a general Bayesian algorithm for constructing the joint distribution of predicted risks and response values based on summary statistics of model performance in previous studies. For statistical metrics of performance, we propose sample size determination rules that either target desired expected precision, or a desired assurance probability that the precision criteria will be satisfied. For NB, we propose rules based on optimality assurance (the probability that the planned study correctly identifies the most beneficial strategy) and the expected value of sample information (EVSI, the expected gain in net benefit from the planned validation study). We showcase these developments in a case study on the validation of a risk prediction model for deterioration of hospitalized COVID-19 patients. Results: The contemporary approach, based on fixed values of c-statistic, O/E ratio, and calibration slope from the development study, with target 95%CI width of, respectively, 0.10, 0.22, and 0.30, would recommend a sample size of 1,056, dictated by the desired precision around the calibration slope. To implement the Bayesian approach and to account for the uncertainty of the model's predictive performance, we constructed predictive distributions for prevalence, c-statistic, mean calibration error, and calibration slope for a new validation study. Targeting the same expected CI width would result in a sample size of 1,070. On the other hand, demanding 90% assurance for meeting these precision criteria would result in a sample size of 1,174. In terms of NB, an optimality assurance of 90% was achieved at a sample size of 306. The EVSI curve demonstrated a diminishing margin with sample sizes >500. Conclusion: Compared to the conventional sample size calculation methods, a Bayesian approach requires explicit quantification of uncertainty around model performance. In exchange it enables various decision rules based on expected value, assurance probability, and value of information. In our case study, EVSI calculations indicated that the precision criterion around calibration slope could potentially be relaxed without much utility loss.

Nahid Sadr, Department of Mathematics and Statistics, Université de Sherbrooke

Contributed Session: Tuesday

Jun 24 15:30-17:00 HC1315

Max-Stability and Sampling Methods for Distorted Extreme Value Copulas

Distorting distributions in the multivariate setting is a useful tool to account for model uncertainty in actuarial science or finance. Specifically, distortions of copulas, which represent the underlying dependence structure, allow us to generate new dependence structures from existing ones, offering flexibility in modeling dependent risks. In this presentation, we investigate the max-domain of attraction problem for Morillas-type distortions of copulas. Furthermore, we establish conditions under which distorted extreme-value copulas retain max-stability. Additionally, multivariate risk measures for generalized extreme value distributions under Morillas-type distortions of the associated dependence structure are discussed. We also propose a simulation algorithm for Morillas-type copula distortions, addressing a gap in the literature and providing a tool for generating distorted dependence structures.

Nathan Phelps, University of Western Ontario

Contributed Session 4: Wednesday

Jun 25 15:30-17:00 HC1325

Platt's scaling for calibration after undersampling—limitations and how to address them

When modelling data where the response is dichotomous and imbalanced, response-based sampling where a subset of the majority class is retained (i.e., undersampling) is often used to create more balanced training datasets prior to modelling. However, models fit to undersampled data, which we refer to as base models, generate biased predictions. There are several calibration methods that can be used to combat this bias, one of which is Platt's scaling. Here, logistic regression is used to model the relationship between the base model's original predictions and the response. Despite its popularity for calibrating models after undersampling, Platt's scaling was not designed for this purpose. Our work presents what we believe is the first detailed study focused on the validity of using Platt's scaling to calibrate models after undersampling. We demonstrate that Platt's scaling is generally not appropriate for calibration in this setting. In particular, Platt's scaling is unable to calibrate a base model that is well-calibrated to its undersampled training dataset, which can lead to severe underestimation of event probabilities for the cases most likely to belong to the minority class. This is especially problematic in fields focused on mitigating the risk of rare events, such as wildfire management.

Nathan Sandholtz, Brigham Young University

Data Science in Sports Analytics

Jun 25 10:00-11:30 HC1700 Labatt Hall

Investigating the Spatial Component of Serving Strategies in Tennis

We conducted an experiment with the Brigham Young University Men's Tennis Team to investigate the spatial component of serving strategy in tennis. Serve data—including precise spatial coordinates of bounce locations—were collected for eight collegiate athletes, with known intended targets for each serve. Using these data, we estimate each player's execution error, defined as the distribution of bounce locations around the intended targets. Because many serves are unobserved due to contact with the net, we explicitly account for censoring when modeling these distributions. The resulting estimates allow us to assess whether players' stated targets align with their actual behavior, as indicated by the centers of the estimated serve distributions. We extend the analysis by estimating player-specific optimal aiming locations for both first and second serves. To account for the two-stage nature of the problem (players are permitted a second serve if they fail to hit the service region on their first serve), we formulate the decision as a two-period Markov decision process (MDP). Solving this MDP requires estimates of the point win probability surface over the service area, which we obtain using methods developed in prior work. We compare the estimated optimal targets to the players' stated targets and discuss cases of alignment and divergence. Statistically, our work provides a fully Bayesian treatment of a two-stage continuous-action decision problem with censored observations and imperfect execution. In addition to other sport applications (e.g., darts, soccer, football, baseball), our methods could be applied in other settings where noise-corrupted decisions are made in continuous action spaces, such as precision military strikes, medical dosing, and autonomous systems.

Nkechi Grace Okoacha, Pan-Atlantic University, Lagos, Nigeria

Contributed Session: Tuesday

Jun 24 15:30-17:00 HC1315

Power-Lindley Generalized Pareto Distribution: A New Approach for Modeling Heavy-Tailed Data

This research introduces the Power-Lindley Generalized Pareto Distribution (PL-GPD), a new statistical model designed to better capture heavy-tailed data, which is often found in finance, climate studies, and other fields where extreme events are significant. Unlike existing models like the Normal and Skew Normal Generalized Pareto Distributions (NGPD and SNGPD) by Debbabi et al. (2012) and Debbabi et al. (2016) respectively, which struggle with the complexities of such data, the PL-GPD combines the Power-Lindley distribution (for moderate values) with the Generalized Pareto distribution (for extreme values).Using Maximum Likelihood Estimation (MLE), the PL-GPD was applied to S&P 500 log return data and compared to NGPD and SNGPD based on Akaike Information Criterion (AIC) and Kolmogorov-Smirnov (K-S) statistic. The PL-GPD showed a superior fit, with lower AIC and higher K-S p-values, indicating improved accuracy in modeling rare but impactful events. This enhanced ability to manage extreme data makes PL-GPD especially useful in risk management for fields like finance and insurance

Olivier Thas, Hasselt University, Belgium

High-Dimensional Data Analysis

Jun 23 15:40-17:20 HC1700 Labatt Hall

Adaptive Large Scale Hypothesis Tests

Adaptive tests are tests that adapt themselves to the data so as to increase the power, but they typically require large sample sizes. In large scale hypothesis testing settings, however, there is a huge amount of data available that can be used for adapting the test so as to increase its power towards interesting alternatives. We present a framework for the construction of adaptive test statistics in this setting. We work out the details for linear rank tests, for which the rank scores are estimated from the data, and we show that they asymptotically result in locally most powerful rank tests (LMPRT). A similar principle is also applied to test statistics that resemble the structure of asymptotic linear estimators; here the parameters of a series expansion of the influence function are estimated from the data. This test statistic is shown to converge to the score test statistic for the testing problem under study. Similarly, a third class of adaptive tests generalises likelihood ratio tests. Under mild conditions we prove the consistency of these tests. In a simulation study we demonstrate the improved sensitivity of our adaptive testing procedures. This includes realistic settings for RNASeq and microbiome studies.

Owen Ward, Simon Fraser University

Bayesian Computational Methods

Jun 23 13:30-15:00 HC1700 Labatt Hall

Scalable Bayesian computation for networks utilising Aggregated Relational Data

Aggregated relational data (ARD) is widely collected in fields such as sociology and has been used to answer important questions about social network structure. This is done by gathering answers to prompts of the form “How many people with trait X do you know?”. An important extension for this data is to consider the task of estimating the underlying network structure, using only this aggregate information.
This talk will provide a brief overview of ARD and how it has been used in important applied problems in social network analysis. I’ll then describe how to leverage this aggregate data only to estimate structure in the underlying unobserved network. Finally, I’ll describe a method and the computational tools required to scale Bayesian community detection to massive networks using ARD, which can uncover structure in a large citation network.

Paul N. Zivich, University of North Carolina at Chapel Hill

Pushing Causal Inference Forward: Blending Machine Learning and Statistical Innovation
Jun 24 10:00-11:30 HC1325

Stabilizing Causal Estimates: Tackling Random Seed Dependence in Machine Learning

Use of machine learning for causal effect estimation has continued to be readily adopted. Machine learning, or data-adaptive algorithms, offer the opportunity to more flexibly model nuisance functions (e.g., propensity scores), which may allow for more reliable causal effect estimates. While promising, machine learning had faced several theoretical challenges to its application. These challenges have largely been addressed through various statistical tools, such as rate double robustness and sample-splitting. However, practical challenges to application remain. One of these challenges is the dependence on the seed of pseudo-random number generators (RNG) at several levels. This dependence on RNG seeds is a threat to reliability and trust in machine learning for causal effect estimation, as results may substantively change under different seeds. In this talk, I will review the role RNG and seeds play in causal effect estimation. To reduce this dependence, I will review and illustrate recent proposals to address this dependence. By using these methods to reduce dependence on the RNG seed, data scientists can be more confident in their causal effect estimates.

Quang Vuong, Core Clinical Science

Bayesian Method in Adaptive Trial Design

Jun 23 13:30-15:00 HC1315

Getting Ready for the Estimands Framework: Simulation-Guided Designs with Illness-Death Models to Explore Trade-Offs in Cancer Trials with Treatment Switching

The ICH E9(R1) addendum is a regulatory document that was recently adopted by the Food and Drug Administration, the European Medicines Agency, Health Canada, and other global regulatory agencies. The addendum calls for use of the estimands framework in planning, designing, and analyzing randomized clinical trials (RCTs). It emphasizes specifying post-randomization (i.e., intercurrent) events and their analytic strategies. As existing sample size formulas typically cannot be applied to trial planning with intercurrent events, simulation-guided designs are often required to optimize trial designs. We conducted a simulation study to explore the trade-offs associated with different analytic strategies for treatment switching, a common intercurrent event in oncology.

Rajitha Senanayake, McMaster University

Contributed Session: Tuesday

Jun 24 15:30-17:00 HC1315

repSpat – A Nonparametric Framework for Repeated Spatial Clustering

Spatial clustering often yields spatially contiguous clusters due to the existence of spatial dependency among nearby observations. However, similar spatial patterns can emerge far apart, forming repeated spatial clusters, which traditional methods like constrained hierarchical clustering fail to identify, they rather treat them as distinct clusters. This results in an excess number of clusters. We propose repSpat, a nonparametric framework designed to detect repeated spatial clusters. repSpat first applies constrained clustering, followed by a multivariate block permutation test based on the maximum mean discrepancy statistic to assess the similarity in the spatial distributions of the identified clusters. If the test indicates significant differences, clusters are re-partitioned to enhance the detection of repeated spatial structures, refining the overall clustering results while preserving spatial dependencies. Through simulations, we evaluate repSpat's ability to detect repeated spatial clusters under varying spatial dependencies, numbers of variables, and sample sizes. We also showcase its application to spatial proteomics data from Triple Negative Breast Cancer patients, identifying repeated tumor microenvironments within tissue.

Renny Doig, Simon Fraser University

Bayesian Computational Methods

Jun 23 13:30-15:00 HC1700 Labatt Hall

PANA-C: A parallel MCMC algorithm for annealed Monte Carlo sampling

Annealing can be a powerful tool in the development of advanced Monte Carlo methods designed for sampling from complicated target distributions. Annealed sequential Monte Carlo (ASMC) is one such method that uses a sequential Monte Carlo sampler to propagate a collection of weighted particles along a sequence of annealed distributions to generate a sample from the target. However, the resampling performed within the algorithm can produce a lower quality sample than is desirable. In this talk we present PANA-C, a novel MCMC algorithm that uses a compound proposal mechanism to combine local exploration and propagation of samples across distributions into a single step. Using this proposal we construct an algorithm that has a similar information propagation mechanism to ASMC, but the absence of resampling and constant refreshing from the reference distribution produce higher quality results. The proposed algorithm also provides a convenient, unbiased estimator of the normalizing constant as well as admitting a highly parallel computing structure. Initial results are presented illustrating the performance of the proposed algorithm, with emphasis on multimodal and non-Gaussian target distributions.

Rhonda Rosychuk, University of Alberta

Causal Inference in Observational Studies

Jun 23 13:30-15:00 HC1325

Comparing Approaches for Clinical Decision Rule Development with a Large Multi-jurisdictional Database

Especially during the coronavirus disease 2019 (COVID-19) pandemic, physicians had to make decisions in an evolving environment with limited evidence in the literature and limited clinical guidelines. Making such decisions requires creating or updating available clinical decision rules and the popular AI (artificial intelligence) systems may not be able to accomplish the task. Motivated by the demand, we explore different approaches to develop a clinical decision rule based on a large network of emergency departments (EDs) that formed in response to the pandemic. This talk will focus on our development of a clinical decision rule to risk-stratify COVID-19 patients at risk of hospital admission or death within 72 hours of discharge from the ED. We consider multi-level regression models, along with regularized regression methods, and propose a dynamic strategy for developing the decision rule. This is joint work with Quinn Forzley and X. Joan Hu.

Richard Yan, Simon Fraser University

Contributed Session: Tuesday

Jun 24 15:30-17:00 HC1315

A Generalized Phase I/II Dose Optimization Trial Design With Multi-Categorical and Multi-Graded Outcomes

Pursuing accurate observations and rational assumptions always drives advances in clinical trial design. In recent years, more trials have begun to collect multi-graded outcomes for more informative analyses. At the same time, assumptions other than the traditional monotonicity relationship have been considered in the dose-efficacy curve to be more realistic. Inspired by these two trends, we propose a phase I/II design that simultaneously considers multi-categorical toxicity and efficacy with multi-graded outcomes, measured as quasi-continuous probability based on prespecified weight matrices of clinical significance. Following keyboard design, our approach aims to screen out overly toxic doses by the toxicity probability intervals and adaptively makes dose escalation or de-escalation decisions by comparing the posterior distributions of dose desirability (utility) among the adjacent levels of the current dose. It helps to more accurately identify the optimal biological dose (OBD) in a non-monotonically increasing dose-efficacy relationship. We also comprehensively present the safety, accuracy and reliability performance through numerical simulations in multiple scenarios and compare the results with several already available designs. The benchmarking results of multiple operating characteristics convincingly support that our design leads in overall performance while ensuring robustness.

Roshni Sahoo, Stanford University

Contributed Session: Monday

Jun 23 15:30-17:00 HC1315

Learning from a Biased Sample

The empirical risk minimization approach to data-driven decision making requires access to training data drawn under the same conditions as those that will be faced when the decision rule is deployed. However, in a number of settings, we may be concerned that our training sample is biased in the sense that some groups (characterized by either observable or unobservable attributes) may be under- or over-represented relative to the general population; and in this setting empirical risk minimization over the training set may fail to yield rules that perform well at deployment. We propose a model of sampling bias called conditional \Gamma-biased sampling, where observed covariates can affect the probability of sample selection arbitrarily much but the amount of unexplained variation in the probability of sample selection is bounded by a constant factor. Applying the distributionally robust optimization framework, we propose a method for learning a decision rule that minimizes the worst-case risk incurred under a family of test distributions that can generate the training distribution under \Gamma-biased sampling. We apply a result of Rockafellar and Uryasev to show that this problem is equivalent to an augmented convex risk minimization problem. We give statistical guarantees for learning a model that is robust to sampling bias via the method of sieves, and propose a deep learning algorithm whose loss function captures our robust learning target. We empirically validate our proposed method in a case study on prediction of mental health scores from health survey data and a case study on ICU length of stay prediction.

Ruitao Lin, MD Anderson Cancer Center

Bayesian Method in Adaptive Trial Design

Jun 23 13:30-15:00 HC1315

TODO: A Triple-Outcome Double-Criterion Optimal Design for Dose Monitoring-and-Optimization in Multi-Dose Randomized Trials

Detecting the efficacy signal and determining the optimal dose are critical steps to increase the probability of success and expedite the drug development in cancer treatment. After identifying a safe dose range through phase I studies, conducting a multi-dose randomized trial becomes an effective approach to achieve this objective. However, there have been limited formal statistical designs for such multi-dose trials, and dose selection in practice is often ad hoc, relying on descriptive statistics. We propose a Bayesian optimal two-stage design to facilitate rigorous dose monitoring and optimization. Utilizing a flexible Bayesian dynamic linear model for the dose-response relationship, we employ dual criteria to assess dose admissibility and desirability. Additionally, we introduce a triple-outcome trial decision procedure to consider dose selection beyond clinical factors. Under the proposed model and decision rules, we develop a systematic calibration algorithm to determine the sample size and Bayesian posterior probability cutoffs to optimize specific design operating characteristics. Furthermore, we demonstrate how to concurrently assess toxicity and efficacy within the proposed framework using a utility-based risk-benefit trade-off. To validate the effectiveness of our design, we conduct extensive simulation studies across a variety of scenarios, demonstrating its robust operating characteristics.

Sankhapali Polgolla, University of Calgary

Contributed Session 1: Monday

Jun 23 15:40-17:20 HC1315

Adjustable Robust Optimization Reformulations for Support Vector Machines

Semiparametric mixture models have been extensively studied in the past decades, due to its flexibility in adapting to different data structures without strong parametric assumptions. Numerous studies have explored semiparametric location-shifted mixture models, establishing their identifiability and various estimation methods. However, for the semiparametric location-scale mixture models, as a natural but broader extension of location-shifted models, investigation remains very limited. To fill the gap, our research focuses on the semiparametric two-component location-scale mixture models. This model consists of five unknown parameters, i.e. mixing proportion, two location parameters and two scale parameters, and one nuisance parameter, which is the symmetric density function with unit standard deviation (for the sake of identifiability) called base function. We first proved the identifiability of this model over the parameter space. Then we applied the minimum distance technique and constructed the minimum Hellinger distance estimators (MHDE) of the unknown parameters after recovering the base function. For this MHDE, we proved its asymptotic properties including its existence, continuity as a functional and consistency . The finite-sample performance of the MHDE illustrated by simulation studies and real data analysis showed that our proposed MHDE is very competitive with widely used methods for normal or light-tailed components, while it performs better than those methods for heavy-tailed components.

Samir Arora, Simon Fraser University

Contributed Session 4: Wednesday

Jun 25 15:30-17:00 HC1325

SPAFIT: Bayesian Empowered Parameter-Efficient Fine-Tuning Search and Model Training for LLMs

Fine-tuning large language models (LLMs) is computationally expensive due to their high parameter density, making it challenging to adapt them to diverse downstream tasks. Parameter-Efficient Fine-Tuning (PEFT) methods have emerged to address this by reducing the number of trainable parameters while maintaining comparable performance to full fine-tuning. However, selecting the most suitable PEFT method for a given use case and model is non-trivial due to the variety of available methods and the associated design choices, such as selecting hyperparameters and determining the layers in which to insert PEFT modules. To address this, we developed an algorithm that leverages Annealed Sequential Monte Carlo (SMC) with data subsampling to optimize fine-tuning across model layers. This approach efficiently computes likelihoods, integrates Metropolis-adjusted Langevin Algorithm (MALA) for updates, and generates samples from the posterior distribution of trainable parameters, enabling predictions via Bayesian Model Averaging (BMA).

Scott Powers, Rice University

Data Science in Sports Analytics

Jun 25 10:00-11:30 HC1700 Labatt Hall

Winning Baseball Games by Solving Statistics Puzzles

We discuss three fun applications of statistical problem solving to answer research questions in baseball. First, we present a twist on supervised learning to predict outcomes from pitch tracking data. How do our methods change when our goal is to evaluate the *pitcher* rather than the *pitch*? Second, we analyze the cat-and-mouse game between pitcher and runner introduced by the new pickoff limit in Major League Baseball (MLB). How can the runner optimally choose their leadoff distance with each successive pickoff attempt? Third, we dig into the groundbreaking swing-by-swing bat tracking recently released to the public by MLB: bat speed and swing length, measured at the point of contact (i.e. the outcome). How can we overcome the fact that the outcome determines the point of measurement?

Shifan Jia, Simon Fraser University

Data Science in Sports Analytics

Jun 25 10:00-11:30 HC1700 Labatt Hall

A New Function-on-Function Regression Model for Rowing Data Analysis

Functional regression is a useful tool in sports analysis because it effectively handles the continuous data collected during races. We propose a function-on-function regression model that predicts a functional response by both a nonlinear dynamic effect of a functional predictor and a linear concurrent effect of another functional predictor. The nonlinear dynamic effect is characterized by taking an integral of a time-dependent two-dimensional smooth surface and the linear concurrent effect is modeled through a time-varying coefficient. The model structure combines the flexibility of nonlinear modeling with the interpretability of the linear concurrent effect. To approximate the two-dimensional surface, we use tensor product basis expansions, and for the time-varying coefficient in the concurrent effect, we employ B-spline expansions. The expansion parameters for each effect are estimated iteratively to account for the mutual dependencies between these two estimated effects. Each iteration of parameter estimation involves solving a penalized least squares problem. The proposed method is applied to rowing data to examine how stroke rate influences speed. Our model directly determines the optimal stroke rates for achieving high speeds based on the estimated smooth surface.

Shouxia Wang, Shanghai University of Finance and Economics

Time Series and Financial Modeling

Jun 24 15:30-17:00 HC1325

High Dimensional Data Assimilation by Optimal Multigrid Ensemble Kalman Filter

The ensemble Kalman Filter, as a fundamental data assimilation approach, has been widely used in many fields of earth science, engineering and beyond. However, the computational and storage cost is very expensive when the state variable is high dimensional accompanied by high resolution observation and physical model in ocean data assimilation. Besides, different sources of the observations have different spatial-temporal resolutions. To assimilate multiscale observations efficiently and to reduce the computational and storage cost, we develop the optimal high dimensional multigrid ensemble Kalman Filter algorithm, which combines the multigrid ensemble Kalman Filter with the optimal selection. The optimal multigrid is selected by minimizing the mean square errors between the analysis state obtained by the multigrid ensemble Kalman Filter and the oracle analysis state by the ensemble Kalman Filter with no multigrid algorithm. Numerical studies on the Lorenz-96 and the Shallow Water Equation models illustrate that the proposed optimal high dimensional multigrid ensemble Kalman Filter algorithm can reduce the computational and storage cost while keeping a comparable accuracy. The proposed method is applied to assimilate sea temperature in the Northwest Pacific, which shows good performance compared with the existing methods.

Shuo Feng, Brown University

Data-Driven Decision Making in Public Health

Jun 23 10:00-11:30 HC1315

Living in a Parallel World? Difference-in-differences for infectious disease outcomes

Researchers frequently employ difference-in-differences (DiD) to study infectious disease policy. DiD assumes that treatment and comparison groups would have moved in parallel in expectation, absent the intervention (``parallel trends assumption"). Our work formalizes often unaddressed epidemiological assumptions required for common DiD specifications, assuming an underlying Susceptible-Infectious-Recovered (SIR) data-generating process, and proposes more robust specifications for infectious disease outcomes. We first demonstrate that popular specifications can encode strict assumptions: DiD modeling incident infections or rates will produce biased treatment effect estimates unless untreated potential outcomes for both groups come from a data-generating process with the same initial infection and transmission rates. Modeling log incidence or growth allows for different initial infection rates under an ``infinite susceptible population" assumption, but invokes conditions on transmission parameters. We propose alternative specifications based on epidemiological parameters -- the effective reproduction number and the effective contact rate -- that are both more robust to differences between groups and can be extended to complex transmission dynamics, such as SEIR models with assortative mixing. We show how treatment effects from these specifications can be transformed to obtain average marginal effects on the incidence scale. In power analyses, we highlight minimal difference between incidence and log incidence models; our alternative specifications have lower power than incidence or log incidence, but higher power than log growth. We illustrate practical implications re-analyzing published studies of COVID-19 mask policies.

Sidi Wu, Fuzhou University

Functional and Longitudinal Modeling

Jun 23 15:30-17:00 HC1325

Neural Networks for Functional Data Analysis

Functional data analysis (FDA) is a statistical discipline that analyzes curves, surfaces and any random variables defined across infinite-dimensional spaces. While numerous efforts have been dedicated to statistical model development in various aspects of FDA, limited attention has been paid to integrating machine learning techniques into FDA. We develop novel methods using neural networks, proposing models that address concerns related to nonlinearity in regression and representation problems regarding functional data. We carefully design a novel functional output layer and a novel functional input layer for neural networks to accommodate both regularly and irregularly spaced functional data. These layers can be integrated in any neural network architectures since they can be backpropagated through. We evaluate the proposed models and demonstrate that our approaches outperform the conventional methods in capturing relationships and reconstructing curves while concurrently maintaining superior smoothing ability and competitive computational efficiency.

Siying Ma, Simon Fraser University

Contributed Session 3: Wednesday

Jun 25 13:30-15:00 HC1325

Data-efficient Operator Learning based on Fundamental Physics Principles

Scientific machine learning models have great potential for accurately solving complex partial differential equations (PDEs). However, current approaches often overlook the explicit preservation of fundamental physical laws embedded within PDEs and thus fail to predict decomposed terms that reflect basic physics. To address this gap, we propose a multi-task learning framework integrating basic terms decomposed from a PDE directly into the training procedure. Through systematic evaluations on reaction-diffusion and Navier-Stokes equations, our method demonstrates significant improvements in data efficiency, enabling the models to achieve comparable root mean square error (RMSE) performance while using datasets that require considerably fewer computational resources. Additionally, we test model robustness and performance under out-of-distribution conditions and validate its performance on real-world scalar flow datasets. Our results confirm that incorporating fundamental PDE information within current scientific machine learning frameworks notably boosts model performance and data efficiency.

Tanya Kovalova, McMaster University

Contributed Session: Monday

Jun 23 15:30-17:00 HC1315

Sharing alpha during interim analysis with multiple primary outcomes.

Introduction: Multiple primary endpoints are often of interest in clinical trials, but there is a lack of literature on interim analysis methods for multiple endpoints, with the US Food and Drug Administration (FDA) stating that this topic is outside of the scope of their Multiple Endpoints guidance. Background: There are two important considerations for identifying interim analysis methods for multiple endpoints. One is the alpha sharing technique defined for the final analysis, and another is which early stopping boundary was chosen for the interim analysis. Purpose: The goal of this Short Communication is to suggest a possible solution for dealing with multiple endpoints at interim analyses, using the example of a real ongoing study X, and to raise the attention of the audience to this topic. Methods: We are proposing sharing alpha between multiple primary endpoints at the interim analyses in the same manner as defined for the final analysis, such that the interim analysis mimics the final one while using different pre-defined levels of alpha. Conclusions: Sharing alpha levels at interim analyses in the same way as at the final analysis is a logical approach, allowing for a formal multiple testing at the interim look, which is an important part of assessing the progress of a clinical trial.

Tao Wang, University of Victoria

Variance Estimation and Statisical Inference

Jun 24 10:00-11:30 HC1700 Labatt Hall

Distributed Mode Learning

We introduce a novel regression methodology leveraging parametric kernel-based mode estimation, specifically designed to handle datasets that exhibit heavy-tailed distributions or contain significant outliers. To effectively tackle computational burdens associated with large-scale data, our approach incorporates distributed statistical learning methods, markedly reducing memory demands while naturally accommodating dataset heterogeneity across distributed environments. By reformulating the local kernel-based objective into an approximate least squares framework, the method efficiently retains compact, summarized statistics from each local computation unit. These compact summaries allow for accurate global estimation with negligible asymptotic loss. Furthermore, we examine shrinkage estimation using a local quadratic approximation scheme and demonstrate that, under an adaptive LASSO framework, the estimator achieves oracle properties. Simulation studies and practical applications to real-world data underscore the superior performance and robustness of our proposed technique in finite-sample scenarios.

Thierry Chekouo Tekougang, University of Minnesota

New developments on survival analysis and variable selection

Jun 24 10:00-11:30 HC1315

A non-parametric Integrative Bayesian Approach for Variable Selection and Prediction

Linear models may not adequately capture complex, nonlinear associations between outcomes and features. Moreover, with advancements in technology, data from multiple platforms are now collected on the same individuals, necessitating methods that effectively integrate multi-platform data. To address these limitations, we propose a nonparametric Bayesian variable selection approach that employs Gaussian process priors to flexibly model the response surface within a model-based data integration framework. The novelties of our method lie in integrating multi-view data using multiple kernels in a Bayesian framework, allowing for simultaneous variable selection for each data type and accurate prediction. The proposed model measures the importance of each data view in predicting clinical outcomes while performing view-specific variable selection. A key feature of our method is that it can also be utilized in accelerated failure time (AFT) models when dealing with censored time-to-event outcomes. We present several simulation studies where we demonstrate the capability of our approach to detect significant variables across various data platforms and to predict outcomes effectively.

Thomas Farrar, Cape Peninsula University of Technology

Variance Estimation and Statisical Inference

Jun 24 10:00-11:30 HC1700 Labatt Hall

A Class of Auxiliary Models for Variance Estimation in Heteroskedastic Linear Models

The best linear unbiased estimator of the linear predictor parameter vector in linear regression in the presence of heteroskedasticity (heterogeneity of error variances) is the weighted least squares estimator. However, since the true weight matrix—the inverse of the diagonal variance-covariance matrix of the random errors—is usually unknown, practitioners tend to employ feasible weighted least squares (FWLS), in which the variances in the weight matrix are replaced by estimates. Often, these estimates are obtained from an auxiliary model in which the response is the vector of squares of ordinary least squares (OLS) residuals. A shortcoming of such models is that they are not built around the true conditional mean of the response.A new class of models is proposed that is based on the true conditional mean of the squared OLS residuals. The individual models differ in their approach to reducing the dimensionality of the parameter vector, which can be done by, for example, modelling the error variances as a function of the predictors, or assuming certain variances to be equal using hierarchical clustering. If the parametrized model is linear, it can be fitted using inequality-constrained least squares or quadratic programming; if nonlinear, using maximum quasi-likelihood estimation. Hyperparameter tuning and feature selection are also incorporated.A Monte Carlo experiment is designed to evaluate the new models empirically using several appropriate metrics. They outperform existing methods under some experimental conditions, and are found to be a promising approach to both estimation and inference in heteroskedastic linear models.

Thomas Thangarajah, University of Waterloo

Contributed Session 4: Wednesday

Jun 25 15:30-17:00 HC1325

Analyzing Team Performance in Professional Sports using Singular Value Decomposition (SVD)

In this presentation, I will share how we extend the applications of singular value decomposition (SVD) to team sports analysis, focusing on the second-order singular value vectors. By moving beyond first-order analysis, we explore how higher-order singular vectors can reveal deeper insights into team performance dynamics. We demonstrate the utility of this approach through specific applications in basketball (NBA) and baseball (MLB), showcasing how SVD can uncover patterns and relationships that traditional metrics may overlook.

Tiantian Yang, University of Idaho

Statistical Genetics, Disease, and Population Modeling

Jun 25 15:30-17:00 HC1315

An Interpretable Graph Neural Network for Disease Classification with Multi-Omics Data

Omics data play crucial roles in exploring disease pathways, forecasting clinical outcomes, and gaining insights for disease classification. However, the significant challenge of dealing with a relatively small number of samples and a large number of features complicates the development of predictive models for omics data analysis, with inherent sparsity in biological networks and the presence of unknown feature interactions adding further complexities. The advent of Graph Neural Networks (GNN) helps alleviate these challenges by incorporating known functional relationships over a graph. However, many existing GNN models utilize graphs either from existing networks or the generated ones alone, which limits model effectiveness. To overcome this restriction, we proposed an innovative GNN model that integrates information from both externally and internally generated feature graphs. We extensively tested the model through simulations and real data applications, confirming its superior performance in classification tasks compared to existing state-of-the-art baseline models. Furthermore, our GNN model can select features with meaningful interpretations in the biomedical context.

Tianyu Guan, York University

Functional and Longitudinal Modeling

Jun 23 15:30-17:00 HC1325

Historical functional linear model with time-varying delay parameter

The historical functional linear model is a powerful tool for analyzing functional data where the current response depends on past values of a functional predictor. In a hydrological study investigating the rainfall-runoff processes, we observed that the delay (or historical time lag) between the predictor and response can vary over time. To capture this dynamic behavior, we extend the historical functional linear model by introducing a time-varying delay parameter. To estimate this parameter, we propose a novel dynamic nested group bridge penalty. By combining this approach with the bivariate Bernstein basis expansion and penalized least squares, our method effectively identifies the time-varying delay and produces a smooth estimate of the bivariate coefficient function simultaneously. We demonstrate the performance of our approach through simulations and apply it to real hydrological data.

Tim Swartz, Simon Fraser University

Data Science in Sports Analytics

Jun 25 10:00-11:30 HC1700 Labatt Hall

Sports Analytics for Pickleball

Pickleball is the fastest growing sport in the US, and has caught on throughout the world. There are now professional pickleball leagues. However, as a relatively new sport, it has not received a lot of attention in the sports analytics literature. This talk introduces some problems in pickleball and work that has been directed to these problems. We will cover assessing whether pickleball is a weak or strong link sport, the impact of wind in pickleball, tournament design and a probabilistic approach for assessing strategy.

Trevor Campbell, University of British Columbia

Bayesian Computational Methods

Jun 23 13:30-15:00 HC1700 Labatt Hall

Asymptotically Exact Variational Inference via Involutive Iterated Random Functions

Most expressive variational families -- such as normalizing flows -- lack practical convergence guarantees, as their theoretical assurances typically hold only at the intractable global optimum. In this work, we present a general recipe for constructing tuning-free, asymptotically exact variational flows from involutive MCMC kernels. The core methodological component is a novel representation of general involutive MCMC kernels as invertible, measure-preserving iterated random function systems, which act as the flow maps of our variational flows. This leads to three new variational families with provable total variation convergence. Our framework resolves key practical limitations of existing variational families with similar guarantees (e.g., MixFlows), while requiring substantially weaker theoretical assumptions. Finally, we demonstrate the competitive performance of our flows across tasks including posterior approximation, Monte Carlo estimates, and normalization constant estimation, outperforming or matching No-U-Turn sampler (NUTS) and black-box normalizing flows.

Vinky Wang, University of British Columbia

Modeling in Natural and Physical Sciences and Engineering

Jun 25 13:30-15:10 HC1315

Extending Hidden Markov Models for Rhythmicity

Recent advancements in sensor technology have enabled high-resolution data collection in free-living environments, allowing detailed study of biological rhythms and their disturbances. Hidden Markov models (HMMs) provide a framework for linking observable data (e.g., body acceleration) to underlying latent states (e.g., rest-activity cycles). However, existing approaches for modelling latent rhythms using HMMs assume fixed, repeating cycles in transition probabilities, limiting their ability to capture changes in rhythmic patterns over time. We propose a flexible model for temporally varying periodic processes by expressing transition probabilities as sums of penalized smooth functions of time. This approach captures variability across cycles while ensuring continuity. We illustrate the utility of our model through the analysis of actigraph data from patients with mood disorders, deriving interpretable insights into disruptions of their circadian rhythms.

Wenqing He, University of Western Ontario

Robust Statistical Methods for Complex Data Challenges

Jun 23 10:00-11:30 HC1700 Labatt Hall

A unified framework of analyzing missing data and variable selection using regularized likelihood

Missing data arise commonly in applications, and various inference methods have been developed under different missing data mechanisms. However the assessment of a feasible missing data mechanism is difficult due to the lack of validation data. The problem is further complicated by the presence of spurious variables in covariates. Focusing on missingness in the response variable, a unified modeling scheme is proposed by utilizing the parametric generalized additive model to characterize various types of missing data processes. Taking the generalized linear model to facilitate the dependence of the response on the associated covariates, the concurrent estimation and variable selection procedures are developed using regularized likelihood. The proposed methods are appealing in their flexibility and generality; they circumvent the need of assuming a particular missing data mechanism that is required by most available methods. Empirical studies demonstrate that the proposed methods result in satisfactory performance in finite sample settings.

Xiaomeng Ju, NYU

Statistical Imaging and Vision

Jun 23 13:30-15:00 HC1520

Bayesian scalar-on-network regression with applications to brain functional connectivity

We present a Bayesian regression framework that analyzes the relationship between brain connectomes and individual traits. Different from many proposals that use vectorized connectivity matrices as regression predictors, our model preserves the Riemannian geometry of these matrices by projecting them to a tangent space. In addition, we exploit the structural information in the projected matrices, finding a data-driven basis that leads to low-dimensional representations and imposing sparsity to the basis for regularization. The proposal yields a parsimonious regression model that offers meaningful interpretations. We demonstrate the performance of our proposal in simulation settings for regression tasks and through a case study with the Human Connectome Project (HCP) resting state fMRI data.

Xiaoping Shi, University of British Columbia, Okanagan

Innovations in Statistical Theory, Design, and Applications

Jun 24 13:30-15:00 HC1325

Approximate inference with exponential tilting densities: theory and applications

A family of exponential tilting density functions (ETD) is presented and compared with energy functions. This ETD family is shown to be associated with the normal density and the log-gamma density by the minimum cross entropy in information theory. In this paper, we show that ETD minimizes the Kulback-Leiber divergence under some moments constraints. Two examples are provided to illustrate how to approximate a baseline density using ETD. In addition, the normalizing constant of the ETD is approximated by three commonly used approximation methods: Gaussian variational approximation (GVA), Laplace approximation (LA), and saddlepoint approximation (SA). It is shown that the normalizing constant obtained by GVA and SA are asymptotically equivalent, and theoretically, both are more accurate than the one obtained by LA. With the availability of the normalizing constant, likelihood-based asymptotic inference can be obtained. To demonstrate the applicability of the proposed method, it is applied to parameter estimation of the Poisson mixed model and Bayesian inference.

Xiaotian Dai, Illinois State University

Novel statistical methods for complex data analysis

Jun 25 10:00-11:30 HC1315

Incorporating gene ontology and disease ontology into Bayesian genomic selection method

Kidney and lung cancers are among the deadliest diseases, each with multiple subtypes. In this talk, we propose a Bayesian genomic selection method focusing on kidney and lung cancer subtypes with gene ontology and disease ontology information. Our aim is to select genes with similar biological functions, facilitating meaningful biological interpretations by leveraging gene ontology as prior biological information. We also propose linking the regression models of different subtypes through a novel prior that incorporates disease ontology information. The disease ontology information regulates the extent of information sharing across different regressions, allowing the method to capture both the homogeneity and heterogeneity of cancer subtypes in the context of high-dimensional molecular profiling data.

Xiong Yi, The State University of New York at Buffalo

Statistical Genetics, Disease, and Population Modeling

Jun 24 13:30-15:00 HC1700 Labatt Hall

Mitigating Bias in Analyzing Privacy-Preserved Survival Data

Sharing time-to-event data is beneficial for enabling collaborative research efforts (e.g., survival studies), facilitating the design of effective interventions, and advancing patient care. However, sharing the exact survival curves poses concerns over privacy. Although there are several popular privacy-protecting solutions (e.g., binning, differential privacy) offer strong protection on the data, the ``sanitized" data usually has low utility and can result in misleading statistical inference. In this work, we first investigate the distortion in bias and variance in regression analysis of sanitized survival data under popular privacy-protecting solutions and provide a strategy to mitigate the bias in estimators with sanitized survival data.

Xuewen Lu, University of Calgary

Advanced Statistical Modeling for Complex Time-to-Event and Spatial-Temperal Data

Jun 24 13:30-15:00 HC1700 Labatt Hall

Variable Selection for the Generalized Odds Rate Non-mixture Cure Model with interval-censored Data

In medical studies, survival times are commonly interval-censored by hospital visits, at the same time, risk factors or covariates may be high-dimensional. Moreover, there exists a cured sub-population, where individuals never experience the event of interest. Under such a situation, it is a challenging task to model the association between interval-censored survival time and a diverging number of risk factors. This paper studies variable selection methods for the generalized odds rate non-mixture cure model with interval-censored data using penalized semiparametric likelihood function when the dimension of the covariates is diverging. The proposed model encompasses the proportional hazards (PH) and proportional odds (PO) non-mixture cure models as special cases. We utilize the broken adaptive bridge (BAR) penalty and other penalties for regularization and study the asymptotic properties of the resultant estimators of regression parameters, including the oracle property and group effect property. To estimate the unknown hazard function, the sieve method based on Bernstein Polynomials is employed. To facilitate the computation, we implement a novel penalized expectation maximization (EM) algorithm. Furthermore, a simulation study is conducted to assess the finite sample performance of the proposed methods and compare the performance of different penalized methods. Finally, the method is applied to a real data set for illustration.

Yidong Zhou, UC Davis

Causal Inference in Observational Studies

Jun 23 13:30-15:00 HC1325

Geodesic Causal Inference

Adjusting for confounding and imbalance when establishing statistical relationships is an increasingly important task, and causal inference methods have emerged as the most popular tool to achieve this. Causal inference has been developed mainly for regression relationships with scalar responses and also for distributional responses. We introduce here a general framework for causal inference when responses reside in general geodesic metric spaces, where we draw on a novel geodesic calculus that facilitates scalar multiplication for geodesics and the quantification of treatment effects through the concept of geodesic average treatment effect. Using ideas from Fréchet regression, we obtain a doubly robust estimation of the geodesic average treatment effect and results on consistency and rates of convergence for the proposed estimators. We also study uncertainty quantification and inference for the treatment effect. Examples and practical implementations include simulations and data illustrations for responses corresponding to compositional responses as encountered for U.S. statewise energy source data, where we study the effect of coal mining, network data corresponding to New York taxi trips, where the effect of the COVID-19 pandemic is of interest, and the studying the effect of Alzheimer's disease on connectivity networks.

Yiming Tang, Shanghai Lixin University of Accounting and Finance

Time Series and Financial Modeling

Jun 24 15:30-17:00 HC1325

A Dual-Basis Multiscale Mixing Architecture for Long-Term Irregular Time Series Forecasting

Irregularly sampled multivariate time series present significant challenges for forecasting tasks, particularly in capturing periodic patterns and long-term trends across varying time scales. In this work, we propose a novel basis-function-based multiscale mixing architecture that explicitly disentangles periodic and trend components in irregular time series. To capture periodic behaviors at multiple frequencies, we utilize Fourier basis functions, while B-spline basis functions are employed to model multi-scale temporal trends. These basis representations are composed into localized temporal patches and combined through a hierarchical mixing strategy, enabling the model to align and fuse information across variables and time resolutions. This design facilitates robust long-term forecasting under severe sampling irregularity, while also providing interpretable decompositions of the temporal structure. Extensive experiments on real-world datasets demonstrate that our method consistently outperforms state-of-the-art baselines in both accuracy and generalizability across diverse irregular time series forecasting tasks.

Ying Yuan, MD Anderson Cancer Center

Bayesian Method in Adaptive Trial Design

Jun 23 13:30-15:00 HC1315

A Bayesian latent-subgroup phase I/II platform design to co-optimize doses in multiple indications

The US Food and Drug Administration (FDA) launched Project Optimus to reform the dose optimization and dose selection paradigm in oncology drug development, calling for the paradigm shift from finding the maximum tolerated dose to the identification of optimal biological dose (OBD). Motivated by a real-world drug development program, we propose a master-protocol-based platform trial design to simultaneously identify OBDs of a new drug, combined with standards of care or other novel agents, in multiple indications. We propose a Bayesian latent subgroup model to accommodate the treatment heterogeneity across indications, and employ Bayesian hierarchical models to borrow information within subgroups. At each interim, we update the subgroup membership and dose-toxicity and -efficacy estimates, as well as the estimate of the utility for risk-benefit tradeoff, based on the observed data across treatment arms to inform the arm-specific decision of dose escalation and de-escalation and identify the optimal biological dose for each arm of a combination partner and an indication. The simulation study shows that the proposed design has desirable operating characteristics, providing a highly flexible and efficient way for dose optimization. The design has great potential to shorten the drug development timeline, save costs by reducing overlapping infrastructure, and speed up regulatory approval.

Yueyang Han, Simon Fraser University

Bayesian Method in Adaptive Trial Design

Jun 23 13:30-15:00 HC1315

Testing the Effectiveness of Treatment for Cancers for which the Endpoint is Survival Using Bayesian Subgroup Analysis

We propose a basket trial design that tests the effectiveness of a new treatment for several types of cancers where the endpoint is the survival time. During the trial conduct, Bayesian subgroup analysis is conducted to classify the cancer types into different clusters according to both the survival time and the longitudinal biomarker measurements of the patient. Finally, we make Bayesian inferences to decide whether to stop recruiting patients for each cluster early and make conclusions about whether the treatment is effective for each cluster according to the estimated median survival time. The simulation study shows that our proposed method performs better than the independent approach and the Bayesian Hierarchical Modeling (BHM) method in most of the scenarios.

Zhenhua Lin, National University of Singapore

Robust Statistical Methods for Complex Data Challenges

Jun 23 10:00-11:40 HC1700 Labatt Hall

Two-sample distribution tests in high dimensions via max-sliced Wasserstein distance and bootstrapping

Two-sample hypothesis testing is a fundamental statistical problem for inference about two populations. In this paper, we construct a novel test statistic to detect high-dimensional distributional differences based on the max-sliced Wasserstein distance to mitigate the curse of dimensionality. By exploiting an intriguing link between the distance and suprema of empirical processes, we develop an effective bootstrapping procedure to approximate the null distribution of the test statistic. One distinctive feature of the proposed test is the ability to construct simultaneous confidence intervals for the max-sliced Wasserstein distances of projected distributions of interest. This enables not only the detection of global distributional differences but also the identification of significantly different marginal distributions between two populations, without the need for additional tests. We establish the convergence of Gaussian and bootstrap approximations of the proposed test, based on which we show that the test is asymptotically valid and powerful as long as the considered max-sliced Wasserstein distance is adequately large. The merits of our approach are illustrated via simulated and real data examples.

Zhou Lan, Brigham and Women's Hospital, Havard Medical School

New developments on survival analysis and variable selection

Jun 24 10:00-11:40 HC1315

Fiber Microstructure Quantile (FMQ) Regression: A Novel Statistical Approach for Analyzing White Matter Bundles from Periphery to Core

The structural connections of the brain's white matter are critical for brain function. Diffusion MRI tractography enables the in-vivo reconstruction of white matter fiber bundles and the study of their relationship to covariates of interest, such as neurobehavioral or clinical factors. In this work, we introduce Fiber Microstructure Quantile (FMQ) Regression, a new statistical approach for studying the association between white matter fiber bundles and scalar factors (e.g., cognitive scores). Our approach analyzes tissue microstructure measures based on quantile-specific bundle regions. These regions are defined in a data-driven fashion according to the quantiles of fractional anisotropy (FA) of a population fiber bundle, which pools all individuals' bundles. The FA quantiles induce a natural subdivision of a fiber bundle, defining regions from the periphery (low FA) to the core (high FA) of the population fiber bundle. To investigate how fiber bundle tissue microstructure relates to covariates of interest, we employ the statistical technique of quantile regression. Unlike ordinary regression, which only models a conditional mean, quantile regression models the conditional quantiles of a response variable. This enables the proposed analysis, where a quantile regression is fitted for each quantile-specific bundle region. To demonstrate FMQ Regression, we perform an illustrative study in a large healthy young adult tractography dataset derived from the Human Connectome Project-Young Adult (HCP-YA), focusing on particular bundles expected to relate to particular aspects of cognition and motor function. In comparison with traditional regression analyses based on FA Mean and Automated Fiber Quantification (AFQ), we find that FMQ Regression provides a superior model fit with the lowest mean squared error. This demonstrates that FMQ Regression captures the relationship between scalar factors and white matter microstructure more effectively than the compared approaches. Our results suggest that FMQ Regression, which enables FA analysis in data-driven regions defined by FA quantiles, is more powerful for detecting brain-behavior associations than AFQ, which enables FA analysis in regions defined along the trajectory of a bundle. FMQ Regression finds significant brain-behavior associations in multiple bundles, including findings unique to males or to females. In both males and females, language performance is significantly associated with FA in the left arcuate fasciculus, with stronger associations in the bundle's periphery. In males only, memory performance is significantly associated with FA in the left uncinate fasciculus, particularly in intermediate regions of the bundle. In females only, motor performance is significantly associated with FA in the left and right corticospinal tracts, with a slightly lower relationship at the bundle periphery and a slightly higher relationship toward the bundle core. No significant relationships are found between executive function and cingulum bundle FA. Our study demonstrates that FMQ Regression is a powerful statistical approach that can provide insight into associations from bundle periphery to bundle core. Our results also identify several brain-behavior relationships unique to males or to females, highlighting the importance of considering sex differences in future research.