International Chinese Statistical Association

泛華統計協會

**Canada Chapter**

ICSA Canada Chapter 2022 Symposium
Statistics: From Data to Knowledge

# Program Book

July 8-10, 2022
Banff Center, Banff, AB, Canada

# Sponsors


Canadian Statistical Sciences Institute
Institut canadien des sciences statistiques

*Data • Discoveries • Decisions*
*Données • Découvertes • Décisions*

Pacific Institute *for the* Mathematical Sciences

# Welcome to the Fifth ICSA-Canada Chapter 2022 Symposium

Welcome to the fifth International Chinese Statistical Association (ICSA) Canada Chapter 2022 Symposium in Banff! The theme of the conference is "Statistics: From Data to Knowledge" and it will feature a rich scientific program focusing on broad areas and the latest developments and innovations in statistics and data sciences and their applications in statistics. It provides a great opportunity and a venue to bring together about 200 statisticians and researchers from Canada and other countries to present and discuss research and practices in-person after more than two years of global COVID-19 lock-down.

The ICSA Canada Chapter was found in 2012. The first biennial symposium was held in Toronto in 2013. The fourth symposium was held in Kingston in 2019. The fifth symposium was scheduled in 2021 and it was postponed to this year because of the pandemic. This year also marks the 10 year anniversary of the Chapter. We will get together to celebrate the $10^{th}$ anniversary and to appreciate all the enthusiastic supporters of the Chapter led by the founding Chair Grace Yi to create this Chapter 10 years ago in Canada. The statistics community in Canada has been growing rapidly, and the Chapter and its biennial symposium soon became one of the best venues for statisticians in Canada and around the world after their inception. The organizing committee of the 2022 symposium has been working hard over the last two years and trying best to make the symposium this year a successful one despite many uncertainties due to pandemic restrictions. More detailed information about the symposium, including abstracts of all invited talks, can be found at https://icsa-canada-chapter.org/symposium2022/.

The symposium venue is Banff Centre for Arts and Creativity, located in Banff, a beautiful resort town within Banff National Park in the province of Alberta. The peaks of Mt. Rundle and Mt. Cascade, part of the Rocky Mountains, dominate its skyline. It is approximately 1.5 hours west of Calgary and can be reached by car or by transit from the Calgary International Airport. Please visit https://www.banffcentre.ca/ for details about Banff Centre, https://banff.ca/ for details about the city of Banff, and https://www.pc.gc.ca/en/pn-np/ab/banff for details about Banff National Park.

The symposium this year is sponsored by the Canadian Statistical Sciences Institute (CANSSI) and The Pacific Institute for the Mathematical Sciences (PIMS). There are also student volunteers from the University of Alberta. We sincerely thank the sponsors and volunteers for their strong supports!

Welcome to Banff and the beautiful Canadian Rocky Mountains!

# ICSA – Canada Chapter Executive Committee

- Joan Hu, Chair-Elect, Simon Fraser University
- Yingwei Peng, Chair, Queen's University
- Liqun Wang, Past-Chair, University of Manitoba
- Leilei Zeng, Secretary/Treasurer, University of Waterloo

# ICSA – Canada Chapter Regional Representatives

- Cindy Feng, Canada East:  Cindy Feng, Dalhousie University
- Juxin Liu, Canada West, University of Saskatchewan
- Sunny Wang, Canada Central, Wilfrid Laurier University

# Symposium Organizing Committee

- Dehan Kong, Chair of the Scientific Program, University of Toronto
- Linglong Kong, Chair of the Local Committee, University of Alberta
- Joan Hu, Chapter Chair-Elect, Simon Fraser University
- Yingwei Peng, Chapter Chair, Queen's University
- Liqun Wang, Chapter Past Chair, University of Manitoba
- Leilei Zeng, Chapter Secretary/Treasurer, University of Waterloo

# Program Committee

- Dehan Kong, University of Toronto (Chair of Program Committee)
- Hua Shen, University of Calgary
- Liangliang Wang, Simon Fraser University
- Yi Yang, Mcgill University
- Yeying Zhu, University of Waterloo

# Local Arrangements Committee

- Linglong Kong, University of Alberta (Chair of Local Committee)
- Bei Jiang, University of Alberta

# Symposium Website

https://icsa-canada-chapter.org/symposium2022/

# Symposium Venue

Symposium activities take place at Max Bell (MB) Building, Banff Centre, Alberta, Canada.

# Travel To Symposium Venue

There are several shuttle services from Calgary to Banff:

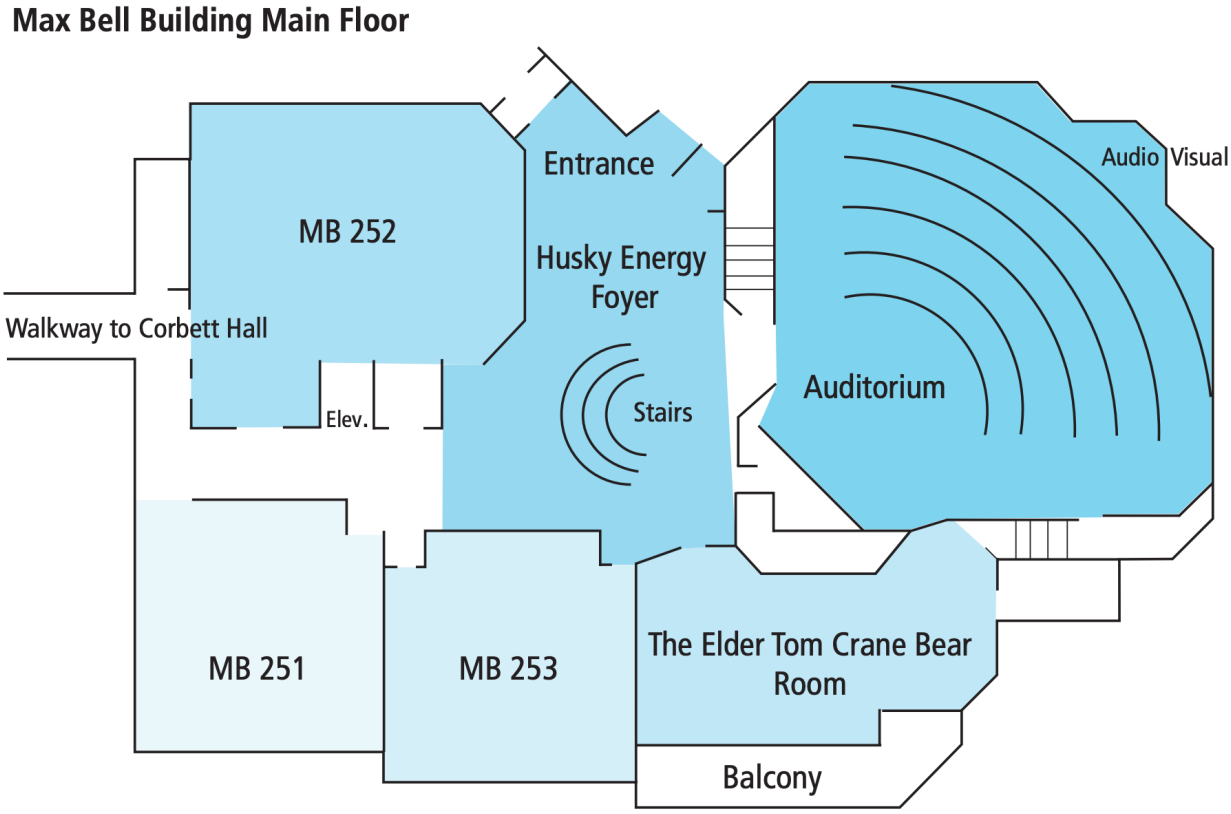
- Banff Airporter.
- Brewster Express.
- Vivo Green.

The BIRS website also has information about the travel to Banff Centre at
https://www.birs.ca/participants/getting-to-birs/

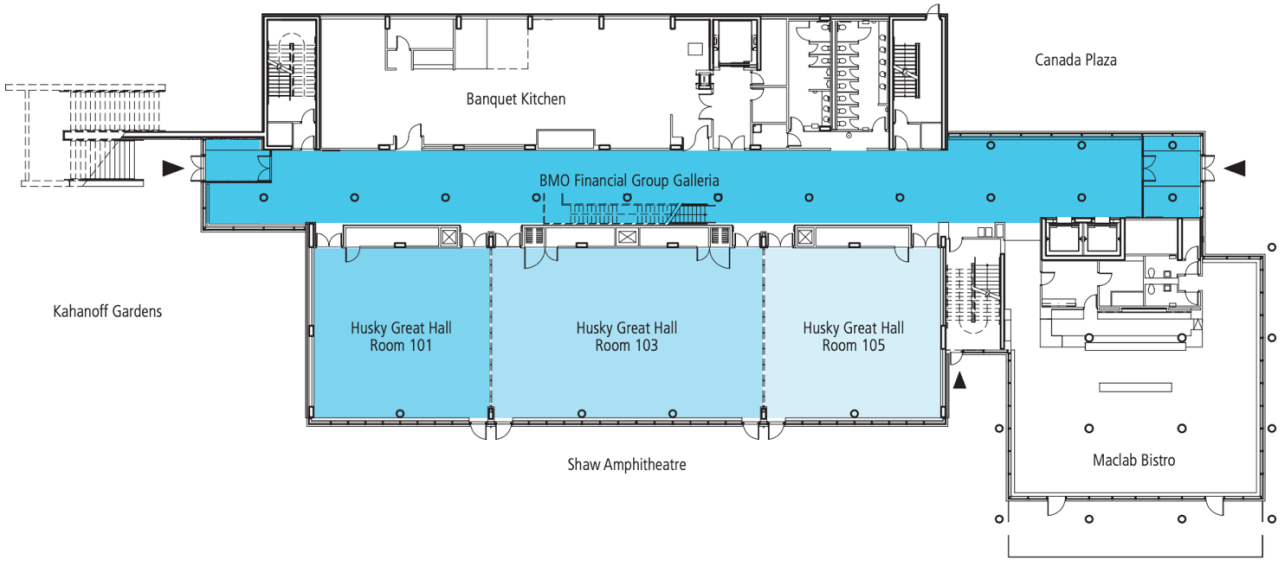
# Rooms and Schedules

- Registration: MB Central Foyer
  - July 7: 5:00pm-9pm,
  - July 8: 8:00am-9pm,
  - July 9: 8:00am-5pm,
  - July 10: 8:00am-12noon.
- Breakfast: Vistas Dining Room, July 8, 9, and 10, 7:00 am – 9:30 am
- Lunch: Vistas Dining Room, July 8, 9, and 10, 11:30 am – 1:30 pm
- Reception: MB Central Foyer, July 8, 6:00 pm – 9:00 pm
- Banquet: Kinnear Centre 103-105, July 9, 7:00 pm – 10:00 pm
- Coffee breaks: MB Central Foyer, July 8, 9, and 10, 9:30 am – 9:50 am & 3:00 pm – 3:20 pm
- Plenary and parallel sessions:
  - MB Auditorium
  - MB 251
  - MB 252
  - MB 253
  - Elder Tom Crane Bear Room

# MB Building Floor Plan



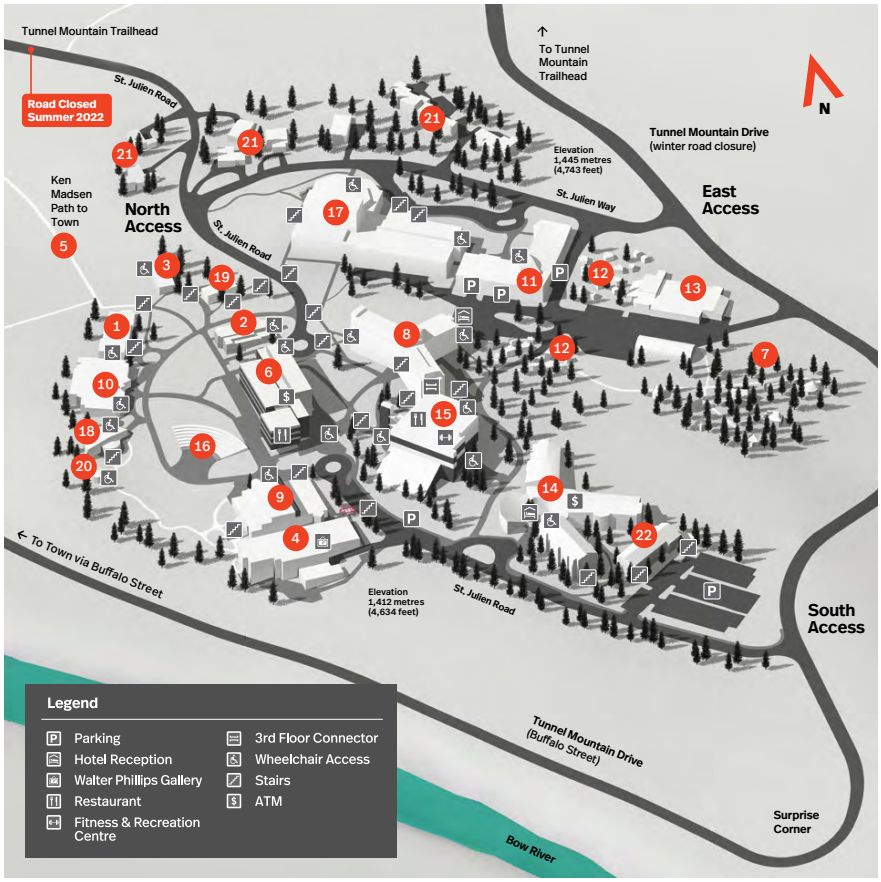

# Kinnear Centre Floor Plan



# Banff Campus Map

## Getting Around Campus

banffcentre.ca | 403.762.6100




BANFF CENTRE FOR ARTS AND CREATIVITY

1 **Corbett Hall**

2 **Donald Cameron Centre**
Administration Offices

3 **Farrally Hall**

4 **Glyde Hall**
Walter Phillips Gallery

5 **Ken Madsen Path to Banff Townsite**

6 **Kinnear Centre for Creativity & Innovation**
Maclab Bistro
Meeting Rooms & Banquets
Paul D. Fleck Library & Archives

7 **Leighton Artists Studios**

8 **Lloyd Hall**
Hotel Reception

9 **Jeanne & Peter Lougheed Building**

10 **Max Bell Building**

11 **Music Building**
Bentley Chamber Music Studio
Rolston Recital Hall

12 **Music Huts**

13 **Physical Facilities Building**
Print Shop
Shipping & Receiving

14 **Professional Development Centre**
Hotel Reception

15 **Sally Borden Building**
Fitness & Recreation Centre
Participant Resources
Three Ravens Restaurant Temporarily Closed
Vistas Dining Room

16 **Shaw Amphitheatre**

17 **Theatre Complex**
Box Office
Jenny Belzberg Theatre
Laszlo Funtek Teaching Wing
Margaret Greenham Theatre
The Club

18 **TransCanada PipeLines Pavilion**
Banff International Research Station

19 **Vinci Hall**

20 **Yurt**

21 **Staff Housing**

22 **Staff Housing**
Becker Hall

# General Schedule

| Time | Function | Location |
|---|---|---|
| **Thursday, July 7** | | |
| 5:00 pm – 9:00 pm | Registration | MB Central Foyer |
| **Friday, July 8** | | |
| 7:00 am – 8:15 am | Breakfast | Vistas Dining Room |
| 8:00 am – 9:00 pm | Registration | MB Central Foyer |
| 8:15 am – 8:30 am | Opening Remarks | MB Auditorium |
| 8:30 am – 9:30 am | Keynote Speech I | MB Auditorium |
| 9:30 am – 9:50 am | Coffee Break | MB Central Foyer |
| 9:50 am – 11:30 am | Parallel Sessions 2-5 | MB (Auditorium, 252, 253), Elder Tom Crane Bear |
| 11:30 am – 1:20 pm | Lunch | Vistas Dining Room |
| 1:20 pm – 3:00 pm | Parallel Sessions 6-10 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 3:00 pm – 3:20 pm | Coffee Break | MB Central Foyer |
| 3:20 pm – 5:00 pm | Parallel Sessions 11-15 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 5:00 pm – 9:00 pm | Poster Session | MB Central Foyer |
| 6:00 pm – 9:00 pm | Reception | MB Central Foyer |
| **Saturday, July 9** | | |
| 7:00 am – 8:30 am | Breakfast | Vistas Dining Room |
| 8:00 am – 5:00 pm | Registration | MB Central Foyer |
| 8:30 am – 9:30 am | Keynote Speech II | MB Auditorium |
| 9:30 am – 9:50 am | Coffee Break | MB Central Foyer |
| 9:50 am – 11:30 am | Parallel Sessions 17-21 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 11:30 am – 1:20 pm | Lunch | Vistas Dining Room |
| 1:20 pm – 3:00 pm | Parallel Sessions 22-26 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 3:00 pm – 3:20 pm | Coffee Break | MB Central Foyer |
| 3:20 pm – 5:00 pm | Parallel Sessions 27-31 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 5:15 pm – 6:45 pm | AGM | Elder Tom Crane Bear |
| 7:00 pm – 10:00 pm | Banquet | Kinnear Centre 103-105 |
| **Sunday, July 10** | | |
| 7:00 am – 8:30 am | Breakfast | Vistas Dining Room |
| 8:00 am – 12 noon | Registration | MB Central Foyer |
| 8:30 am – 9:30 am | Keynote Speech III | MB Auditorium |
| 9:30 am – 9:50 am | Coffee Break | MB Central Foyer |
| 9:50 am – 11:30 am | Parallel Sessions 33-37 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 11:30 am – 1:20 pm | Lunch | Vistas Dining Room |
| 1:20 pm – 3:00 pm | Parallel Sessions 38-41 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |
| 3:00 pm – 3:20 pm | Coffee Break | MB Central Foyer |
| 3:20 pm – 5:00 pm | Parallel Sessions 42-46 | MB (Auditorium, 251, 252, 253), Elder Tom Crane Bear |

# Schedule for July 8, 9, 10

### 8:15am-8:30am, Opening Remarks, MB Auditorium

## Plenary Talk I
## 8:30am - 9:30am, Friday, July 8th

**Session** 1: ***Keynote Speech 1***
Organizer and Chair: Yingwei Peng, Queen's University
**Room: MB Auditorium, Time: 8:30 AM – 9:30 AM**

(1) Xiaotong Shen, University of Minnesota
Title: *Data Perturbation*
Abstract: Data perturbation is a technique for generating synthetic data by adding "noise" to original data, which has a wide range of applications, primarily in data security. Yet, it has not received much attention within data science. In this presentation, I will describe a fundamental principle of data perturbation that preserves the distributional information, thus ascertaining the validity of the downstream analysis and a machine learning task while protecting data privacy. Applying this principle, we derive a scheme to allow a user to perturb data nonlinearly while meeting the requirements of differential privacy and statistical analysis. It yields credible statistical analysis and high predictive accuracy of a machine learning task. Finally, I will highlight multiple facets of data perturbation through examples. This work is joint with B Xuan and R Shen. multiple facets of data perturbation through examples.

### Coffee Break, MB Central Foyer

## Parallel Sessions B
## 9:50am - 11:30pm, Friday, July 8th

**Session** 2: ***Recent Advances in Statistical Learning***
Organizer: Dehan Kong, University of Toronto
Chair: Xinyi Zhang, University of Toronto
**Room: MB Auditorium, Time: 9:50 AM – 11:30 AM**

(1) Weibin Mo, Purdue University
Title: *Learning Optimal Distributionally Robust Individualized Treatment Rules*
Abstract: Recent development in the data-driven decision science has seen great advances in individualized decision making. Given data with individual covariates, treatment assignments and outcomes, policy makers best individualized treatment rule (ITR) that maximizes the expected outcome, known as the value function. Many existing methods assume that the training and testing distributions are the same. However, the estimated optimal ITR may have poor

generalizability when the training and testing distributions are not identical. In this work, we consider the problem of finding an optimal ITR from a restricted ITR class where there are some unknown covariate changes between the training and testing distributions. We propose a novel distributionally robust ITR (DR-ITR) framework that maximizes the worst-case value function across the values under a set of underlying distributions that are "close" to the training distribution. The resulting DR-ITR can guarantee the performance among all such distributions reasonably well. We further propose a calibrating procedure that tunes the DR-ITR adaptively to a small amount of calibration data from a target population. In this way, the calibrated DR-ITR can be shown to enjoy better generalizability than the standard ITR based on our numerical studies.

(2) Ji Zhu, University of Michigan
Title: *Population-Level Balance in Signed Networks*

Abstract: Statistical network models are useful for understanding the underlying formation mechanism and characteristics of complex networks. However, statistical models for signed networks have been largely unexplored. In signed networks, there exist both positive (e.g., like, trust) and negative (e.g., dislike, distrust) edges, which are commonly seen in real-world scenarios. The positive and negative edges in signed networks lead to unique structural patterns, which pose challenges for statistical modeling. In this paper, we introduce a statistically principled latent space approach for modeling signed networks and accommodating the well-known balance theory, i.e., "the enemy of my enemy is my friend" and "the friend of my friend is my friend". The proposed approach treats both edges and their signs as random variables, and characterizes the balance theory with a novel and natural notion of population-level balance. This approach guides us towards building a class of balanced inner-product models, and towards developing scalable algorithms via projected gradient descent to estimate the latent variables. We also establish non-asymptotic error rates for the estimates, which are further verified through simulation studies. We also apply the proposed approach to an international relation network, which provides an informative and interpretable model-based visualization of countries during World War II.

(3) Yingying Fan, University of Southern California
Title: *Asymptotic Properties of High-Dimensional Random Forests*

Abstract: As a flexible nonparametric learning tool, random forests algorithm has been widely applied to various real applications with appealing empirical performance, even in the presence of high-dimensional feature space. Unveiling the underlying mechanisms has led to some important recent theoretical results on the consistency of the random forests algorithm and its variants. However, to our knowledge, all existing works concerning random forests consistency in high dimensional setting were established for various modified random forests models where the splitting rules are independent of the response. In light of this, in this paper we derive the consistency rates for the random forests algorithm associated with the sample CART splitting criterion, which is the one used in the

original version of the algorithm (Breiman2001), in a general high-dimensional nonparametric regression setting through a bias-variance decomposition analysis. Our new theoretical results show that random forests can indeed adapt to high dimensionality and allow for discontinuous regression function. Our bias analysis characterizes explicitly how the random forests bias depends on the sample size, tree height, and column subsampling parameter. Some limitations on our current results are also discussed.

(4) Jinchi Lv, University of Southern California
Title: *High-Dimensional Knockoffs Inference for Time Series Data*
Abstract: The recently introduced framework of model-X knockoffs provides a flexible tool for exact finite-sample false discovery rate (FDR) control in variable selection in arbitrary dimensions without assuming any dependence structure of the response on covariates. It also completely bypasses the use of conventional p-values, making it especially appealing in high-dimensional nonlinear models. Existing works have focused on the setting of independent and identically distributed observations. Yet time series data is prevalent in practical applications in various fields such as economics and social sciences. This motivates the study of model-X knockoffs inference for time series data. In this paper, we make some initial attempt to establish the theoretical and methodological foundation for the model-X knockoffs inference for time series data. We suggest the method of time series knockoffs inference (TSKI) by exploiting the idea of subsampling to alleviate the difficulty caused by the serial dependence. We establish sufficient conditions under which the original model-X knockoffs inference combined with subsampling still achieves the asymptotic FDR control. Our technical analysis reveals the exact effect of serial dependence on the FDR control. To alleviate the practical concern on the power loss because of reduced sample size cause by subsampling, we exploit the idea of knockoffs with copies and multiple knockoffs. Under fairly general time series model settings, we show that the FDR remains to be controlled asymptotically. To theoretically justify the power of TSKI, we further suggest the new knockoff statistic, the backward elimination ranking (BE) statistic, and show that it enjoys both the sure screening property and controlled FDR in the linear time series model setting. The theoretical results and appealing finite-sample performance of the suggested TSKI method coupled with the BE are illustrated with several simulation examples and an economic inflation forecasting application. This is a joint work with Chien-Ming Chi, Yingying Fan and Ching-Kang Ing.

**Session** 3: ***Recent Development in Statistical Computing and Methodology***
Organizer and Chair: Yi Yang, Mcgill University
**Room: MB252, Time: 9:50 AM – 11:30 AM**

(1) Kaiqiong Zhao, University of Alberta
Title: *A Sparse High-Dimensional Generalized Varying Coefficient Model for Identifying Genetic Variants Associated with Regional Methylation Levels*
Abstract: Varying coefficient models offer the flexibility to learn the dynamic

changes of regression coefficients. Despite their good interpretability and diverse applications, in high-dimensional settings, existing estimation methods for such models have important limitations. For example, we routinely encounter the need for variable selection when faced with a large collection of covariates with nonlinear/varying effects on outcomes, and no ideal solutions exist. One illustration of this situation could be identifying a subset of genetic variants with local influence on methylation levels in a regulatory region. To address this problem, we propose a composite sparse penalty that encourages both sparsity and smoothness for the varying coefficients. We present an efficient proximal gradient descent algorithm to obtain the penalized estimation of the varying regression coefficients in the model. A comprehensive simulation study has been conducted to evaluate the performance of our approach in terms of estimation, prediction and selection accuracy. We show that the inclusion of smoothness control yields much better results than having the sparsity-regularization only.

(2) Liangyuan Hu, Rutgers University
Title: *A Flexible Approach for Causal Inference with Multiple Treatments and Clustered Survival Outcomes*

Abstract: When drawing causal inferences about the effects of multiple treatments on clustered survival outcomes using observational data, we need to address implications of the multilevel data structure, multiple treatments, censoring and unmeasured confounding for causal analyses. Few off-the-shelf causal inference tools are available to simultaneously tackle these issues. We develop a flexible random-intercept accelerated failure time model, in which we use Bayesian additive regression trees to capture arbitrarily complex relationships between censored survival times and pre-treatment covariates and use the random intercepts to capture cluster-specific main effects. We develop an efficient Markov chain Monte Carlo algorithm to draw posterior inferences about the population survival effects of multiple treatments and examine the variability in cluster-level effects. We further propose an interpretable sensitivity analysis approach to evaluate the sensitivity of drawn causal inferences about treatment effect to the potential magnitude of departure from the causal assumption of no unmeasured confounding. Expansive simulations empirically validate and demonstrate good practical operating characteristics of our proposed methods. Applying the proposed methods to a dataset on older high-risk localized prostate cancer patients drawn from the National Cancer Database, we evaluate the comparative effects of three treatment approaches on patient survival, and assess the ramifications of potential unmeasured confounding. The methods developed in this work are readily available in the package riAFTBART.

(3) Ying Zhou, University of Toronto
Title: *The Promises of Parallel Outcomes*

Abstract: A key challenge in causal inference from observational studies is the identification of causal effects in the presence of unmeasured confounding. In this paper, we introduce a novel framework that leverages information in multiple parallel outcomes for causal identification with unmeasured confounding.

Under a conditional independence structure among multiple parallel outcomes, we achieve nonparametric identification of causal effects with at least three parallel outcomes. Our identification results pave the road for causal effect estimation with multiple outcomes. In the Supplementary Material, we illustrate the promises of this framework by developing nonparametric estimating procedures in the discrete case, and evaluating their finite sample performance through numerical studies.

(4) Shu Yang, North Carolina State University
Title: *Generalizable Survival Analysis of Randomized Clinical Trials with Observational Studies*

Abstract: In the presence of heterogeneity between the randomized controlled trial (RCT) participants and the target population, evaluating the treatment effect solely based on the RCT often leads to biased quantification of the real-world treatment effect. To address the problem of lack of generalizability for the treatment effect estimated by the RCT sample, we leverage observational studies with large samples that are representative of the target population. This paper concerns evaluating treatment effects on survival outcomes for a target population and considers a broad class of estimands that are functionals of treatment-specific survival functions, including differences in survival probability and restricted mean survival times. Motivated by two intuitive but distinct approaches, i.e., imputation based on survival outcome regression and weighting based on inverse probability of sampling, censoring, and treatment assignment, we propose a semiparametric estimator through the guidance of the efficient influence function. The proposed estimator is doubly robust in the sense that it is consistent for the target population estimands if either the survival model or the weighting model is correctly specified, and is locally efficient when both are correct. In addition, as an alternative to parametric estimation, we employ the nonparametric method of sieves for flexible and robust estimation of the nuisance functions and show that the resulting estimator retains the root-$n$ consistency and efficiency, the so-called rate-double robustness. Simulation studies confirm the theoretical properties of the proposed estimator and show it outperforms competitors. We apply the proposed method to estimate the effect of adjuvant chemotherapy on survival in patients with early-stage resected non-small lung cancer.

**Session** 4: ***Statistics and Economics in Data Science***
Organizer: Linglong Kong, University of Alberta
Chair: Yafei Wang, University of Alberta
**Room: Elder Tom Crane Bear, Time: 9:50 AM – 11:30 AM**

(1) Matias Salibian Barrera, The University of British Columbia
Title: *Functional Principal Components for Sparse Longitudinal Data*

Abstract: We propose a new method to perform functional principal components analysis (FPCA) that can be applied to longitudinal data with few observations per trajectory and which relies on relatively weak regularity assumptions. We use local regression to estimate the values of the covariance function taking advantage

of the fact that for elliptically distributed random vectors the conditional location parameter of some of its components given others is a linear function of the conditioning set. Furthermore, this approach can naturally be modified to obtain robust functional principal component estimators. Numerical experiments show that this method compares favourably to existing alternatives in the literature.

(2) Michal Pesta, Charles University
Title: *Infinitely Stochastic Micro Forecasting*

Abstract: Stochastic forecasting and risk valuation are now front burners in a list of applied and theoretical sciences. In this work, we propose an unconventional tool for stochastic prediction of future expenses based on the individual (micro) developments of recorded events. Considering a firm, enterprise, institution, or any entity, which possesses knowledge about particular historical events, there might be a whole series of several related subevents: payments or losses spread over time. This all leads to an infinitely stochastic process at the end. The aim, therefore, lies in predicting future subevent flows coming from already reported, occurred but not reported, and yet not occurred events. The emerging forecasting methodology involves marked time-varying Hawkes process with marks being other time-varying Hawkes processes. The estimated parameters of the model are proved to be consistent and asymptotically normal under simple and easily verifiable assumptions. The empirical properties are investigated through a simulation study. In the practical part of our exploration, we elaborate a specific actuarial application for micro claims reserving.

(3) Ivan Mizera, University of Alberta
Title: *Functional Profile Techniques for Claims Reserving*

Abstract: One of the most fundamental tasks in non-life insurance, done on regular basis, is risk reserving assessment analysis, which amounts to predict stochastically the overall loss reserves to cover possible claims. The most common reserving methods are based on different parametric approaches using aggregated data structured in the run-off triangles. We propose a rather nonparametric approach, which handles the underlying loss development triangles as functional profiles and predicts the claim reserve distribution through permutation bootstrap. Three competitive functional-based reserving techniques, each with slightly different scope, are presented; their theoretical and practical advantages - in particular, effortless implementation, robustness against outliers, and wide-range applicability - are discussed. Apart from theoretical justifications of the methods, an evaluation of the empirical performance of the designed methods and a full-scale comparison with standard (parametric) reserving techniques are carried on several hundreds of real run-off triangles against the known real loss outcomes. An important objective is also to promote the idea of natural usefulness of the functional reserving methods among the reserving practitioners. (Joint work with Matúš Maciak and Michal Pešta, Charles University, Prague.)

(4) Matus Maciak, Charles University
Title: *Online Regime Switching in a Nonlinear Expectile Model*

Abstract: Regime switching in advanced stochastic models attracts a lot of interest over the last years with many different strategies being proposed in this direction while many complex problems remain still unresolved. We introduce a complex online changepoint detection procedure based on conditional expectile estimation. Nonlinearity of the underlying model improves the overall flexibility, the conditional expectiles—well-known in econometrics for being the only coherent and elicitable risk measure—bring in some additional robustness, and the proposed changepoint detection test is proved to be consistent while the asymptotic distribution of the test statistic under the null hypothesis depends neither on the functional form of the underlying model nor the unknown parameters. This ensure relatively simple and straightforward applicability for real-life situations. Important theoretical details are summarized and finite sample empirical properties are presented.

**Session** 5: ***New Statistical Methods for Modeling Complex Data***
Organizer: Xuewen Lu, University of Calgary
Chair: Liqun Wang, University of Manitoba
**Room: MB253, Time: 9:50 AM – 11:30 AM**

(1) Xiaoke Zhang, George Washington University
Title: *Proximal Learning for Individualized Treatment Regimes under Unmeasured Confounding*

Abstract: Data-driven individualized decision making has recently received increasing research interest. Most existing methods rely on the assumption of no unmeasured confounding, which unfortunately cannot be ensured in practice, especially in observational studies. Motivated by the recently proposed proximal causal inference, we develop several proximal learning approaches to estimating optimal individualized treatment regimes (ITRs) in the presence of unmeasured confounding. In particular, we establish several identification results for different classes of ITRs, exhibiting the trade-off between the risk of making untestable assumptions and the value function improvement in decision making. Based on these results, we propose several classification-based approaches to finding a variety of restricted in-class optimal ITRs and developing their theoretical properties. The appealing numerical performance of our proposed methods is demonstrated via an extensive simulation study and a real data application.

(2) Thierry Chekouo, University of Calgary
Title: *A Bayesian Group Selection with Compositional Responses for Analysis of Radiologic Tumor Proportions and their Genomic Determinants*

Abstract: Volumetric imaging features are used in cancer research to determine the size and the composition of a tumor, and have been shown to be prognostic of overall survival. In this paper, we focus on the analysis of tumor component proportions of brain cancer patients collected through The Cancer Genome Atlas (TCGA) project. Our main goal is to identify pathways and corresponding genes that can explain the heterogeneity of the composition of a brain tumor. In

particular, we focus on the glioblastoma multiform (GBM), as it is the most common malignant brain neoplasm, accounting for 23

(3) Fatemeh Mahmoudi, University of Calgary
Title: *Variable Selection for Semi-Competing Risks Data with Broken Adaptive Ridge Regression*
Abstract: Semi-competing risks data arise when both non-terminal and terminal events are considered in a model. In this framework, terminal event can censor the non-terminal event, but not vice versa. It is known that variable selection is practical in identifying significant risk factors in high-dimensional data. While some recent works on penalized variable selection deal with these competing risks separately without incorporating possible dependence between them, we perform variable selection in an illness-death model using shared frailty where semiparametric hazard regression models are used to model the effect of covariates. We propose a broken adaptive ridge (BAR) penalty to encourage sparsity and conduct extensive simulation studies to compare its performance with other methods. The grouping effect as well as the oracle property of the proposed BAR procedure are also investigated using simulation studies. The proposed method is then applied to the real-life data arising from a Colon Cancer study.

(4) Junhao Zhu, University of Toronto
Title: *Laplacian Optimal Transport Based Reconstruction of Spatial Gene Expression*
Abstract: The spatial expression pattern of cells is vital for inferring the heterogeneity of cells' fate in complex tissue and understanding the tissue function. Although new experimental approaches have been applied to sequencing RNA at the single-cell resolution within the context of the tissues, it provides limited resolutions for expressions since only very few marker genes among thousands of genes are measured. Here, we introduce a method based on linear-model and Laplacian Optimal Transport to integrate spatial reference data and scRNA-seq data to study spatial courses of cells within a tissue. We apply the method to Drosophila scRNA-seq data and successfully reconstruct spatial gene-expression profiles in Drosophila early embryos. The results demonstrate the ability of our approach to provide a biologically interpretable framework for inferences and reconstructions about the spatial expression patterns of cells.

**Lunch, 11:30 AM - 1:20 PM, Vistas Dining Room**

# Parallel Sessions C
# 1:20pm - 3:00pm, Friday, July 8th

**Session** 6: ***Functional and Complex Data Analysis***
Organizer and Chair: Zhenhua Lin, National University of Singapore
**Room: MB Auditorium, Time: 1:20 PM – 3:00 PM**

(1) Zhenhua Lin, National University of Singapore
Title: *High-Dimensional MANOVA via Bootstrapping and Its Application to*

*Functional Data*

Abstract: Presented is a new approach to the problem of high-dimensional multivariate ANOVA via bootstrapping max statistics that involve the differences of sample mean vectors. The proposed method proceeds via the construction of simultaneous confidence regions for the differences of population mean vectors. It is suited to simultaneously test the equality of several pairs of mean vectors of potentially more than two populations. By exploiting the variance decay property that is a natural feature in relevant applications such as functional data analysis, it is possible to achieve dimension-free and nearly-parametric convergence rates for Gaussian approximation, bootstrap approximation, and the size of the test. The proposed methodology, demonstrated with ANOVA problems for functional data and sparse count data, is shown to work well in simulations and real data applications.

(2) Shu Jiang, Washington University
Title: *Predicting Long-Term Breast Cancer Risk with Mammogram Imaging Data*
Abstract: Screening mammography aims to identify breast cancer early and secondarily measures breast density to classify women at higher or lower than average risk for future breast cancer in the general population. Our primary goal in this study is to extract mammogram-based features that augment the well-established breast cancer risk factors to improve prediction accuracy. In this talk, I will present a novel supervised functional principal component analysis to extract image-based features that are ordered by association with the failure times.

(3) Thorsten Koch, ZIB / TU-Berlin
Title: *Dealing with Superhuman Complexity in Data Errors*
Abstract: When conducting projects with industry using highly connected data, we encountered several cases where our analysis detected errors, which were too complex for humans to understand. Examples are irreducible infeasible subsystems (IIS) of large mixed-integer programs (MIP) or bottlenecks in non-linear networks, e.g., pressure coupled pipeline networks. While detecting this kind of problem is a significant achievement, it is also challenging to explain the problems to the practitioners, and removal such errors is extremely difficult. This presentation will show examples, explain the difficulties, and report possibilities on how to deal with the situation. However, significant research in this area is still needed to make complex analyses useful in industrial practice.

(4) Ying Chen, National University of Singapore
Title: *Policy Effectiveness on the Global COVID-19 Pandemic and Unemployment Outcomes: A Large Mixed Frequency Spatial Approach*
Abstract: We propose a mixed frequency spatial VAR (MF-SVAR) modeling framework to measure the effectiveness of policies conditional on the spillover and diffusion effects of the global pandemic and unemployment. We study the effects of two aspects of policy effectiveness, namely policy start date and policy timeliness, from a spatio-temporal perspective. The spatial panel data contain

weekly new case growth rates and monthly unemployment rate changes for 68 countries across six continents at mixed frequencies from January 2020 to August 2021. We find that government policies have a significant impact on the growth of new cases, but only a marginal effect on the change in unemployment rates. A policy's start date is critical for its effectiveness. In terms of both immediate impact on the near term and total impact over the following four weeks, starting a policy in the 4th week of a month is most effective at reducing the growth of new cases. At the same time, starting in the 2nd or 3rd week is counterproductive for a one-time policy start date. In addition, our estimates suggest that the spillover and diffusion effects are much stronger than a country's temporal effect during a global pandemic, both for new case growth and changes in unemployment. We also find that new case growth influences changes in unemployment, but not vice versa. Counterfactual experiments provide further evidence of policy effectiveness in various scenarios and also reveal the main risk-vulnerable and risk-spillover countries. This is a joint work with Xiaoyi Han, Yanli Zhu and Yijiong Zhang. Paper is available at `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4049509`.

**Session** 7: ***Statistical Learning for Complex Data Structures***
Organizer and Chair: Yi Yang, Mcgill University
**Room: MB252, Time: 1:20 PM – 3:00 PM**

(1) Teng Zhang, University of Central Florida
Title: *Alternating Minimization Algorithm for Clustering Mixture Multilayer Network*

Abstract: The paper considers a Mixture Multilayer Stochastic Block Model (MMLSBM), where layers can be partitioned into groups of similar networks, and networks in each group are equipped with a distinct Stochastic Block Model. The goal is to partition the multilayer network into clusters of similar layers, and to identify communities in those layers. Jing et al. (2020) introduced the MMLSBM and developed a clustering methodology, TWIST, based on regularized tensor decomposition. The present paper proposes a different technique, an alternating minimization algorithm (ALMA), that aims at simultaneous recovery of the layer partition, together with estimation of the matrices of connection probabilities of the distinct layers. Compared to TWIST, ALMA achieves higher accuracy both theoretically and numerically.

(2) Ning Hao, University of Arizona
Title: *Quadratic Discriminant Analysis by Projection*

Abstract: Discriminant analysis, including linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA), is a popular approach to classification problems. It is well known that LDA is suboptimal to analyze heteroscedastic data, for which QDA would be an ideal tool. However, QDA is less helpful when the number of features in a data set is moderate or high, and LDA and its variants often perform better due to their robustness against dimensionality. In this talk, we will introduce a new dimension reduction and classifica-

tion method based on QDA. In particular, we define and estimate the optimal one-dimensional (1D) subspace for QDA, which is a novel hybrid approach to discriminant analysis. The new method can handle data heteroscedasticity with number of parameters equal to that of LDA. Therefore, it is more stable than the standard QDA and works well for data in moderate dimensions. We show an estimation consistency property of our method, and compare it with LDA, QDA, regularized discriminant analysis (RDA) and a few other competitors by simulated and real data examples.

(3) Wen Zhou, Colorado State University
Title: *Integrative Group Factor Model for Variable Clustering on Temporally Dependent Date: Optimality and Algorithm*

Abstract: Clustering a large number of variables is fast emerging in a variety of areas, and has become a fundamental problem in statistics and machine learning. Though many algorithmic approaches scatter across the literature, their interpretation is limited and the outputs usually lack guarantees. Furthermore, their explicit and implicit assumptions such as the independence of data and the well-separation between clusters are rather restricted if not unrealistic. In this work, we take the view of model-based clustering, in which the population level clusters are clearly interpreted statistically, to cluster a larger number of variables. The proposed integrative group factor model (iGFM) is compatible with temporally dependent data and allows connections across the variable clusters. In this model, two types of latent factors, the common and unique factors are introduced to model the cross-cluster connection and the within-cluster similarity among variables. We quantify the difficulty of clustering variables based on the iGFM in terms of a permutation-invariant clustering risk and derive the minimax signal threshold, below which no algorithms can cluster variables successfully. Such a threshold is driven by the competition between common and unique factors in the model and does nor request the well-separation of clusters to guarantee a perfect recovery. Based on the spectral decomposition and the idea of linear search, we develop a fast and minimax-optimal algorithm to cluster variables. An interesting phase transition of the clustering performance has been discovered, for which the model parameter space is partitioned into three regions corresponding to cases of impossible to cluster perfectly, possible with guarantees on the optimality, and possible with no provable guarantees, respectively. In addition, we compare our method with another popular model-based method, the G-block model and associated COD algorithm. Extensive simulation studies, as well as careful data analyses on the macroeconomics index data, confirm the advantage of our approach. Finally, we also discuss how to characterize the unknown number of clusters and the extension of our method with a divergent number of clusters.

(4) Mingqi Wu, McGill University
Title: *How Rotational Invariance of Common Kernels Prevents Generalization in High Dimensions*

Abstract: Kernel ridge regression is well-known to achieve minimax optimal rates

in low-dimensional settings. However, its behavior in high dimensions is much less understood. Recent work establishes consistency for high-dimensional kernel regression for a number of specific assumptions on the data distribution. In this paper, we show that in high dimensions, the rotational invariance property of commonly studied kernels (such as RBF, inner product kernels, and fully-connected NTK of any depth) leads to inconsistent estimation unless the ground truth is a low-degree polynomial. Our lower bound on the generalization error holds for a wide range of distributions and kernels with different eigenvalue decays. This lower bound suggests that consistency results for kernel ridge regression in high dimensions generally require a more refined analysis that depends on the structure of the kernel beyond its eigenvalue decay.

**Session** 8: ***Recent Advances in Causal Inference and Missing Data Analysis***
Organizer and Chair: Shu Yang, North Carolina State University
**Room: MB253, Time: 1:20 PM – 3:00 PM**

(1) Jiaying Gu, University of Toronto
Title: *Partial Identification in Nonseparable Binary Response Models with Endogenous Regressors*

Abstract: This paper considers (partial) identification of a variety of counterfactual parameters in binary response models with possibly endogenous regressors. Our framework allows for nonseparable index functions with multi-dimensional latent variables, and does not require parametric distributional assumptions. We leverage results on hyperplane arrangements and cell enumeration from the literature on computational geometry in order to provide a tractable means of computing the identified set. We demonstrate how various functional form, independence, and monotonicity assumptions can be imposed as constraints in our optimization procedure to tighten the identified set. Finally, we apply our method to study the effects of health insurance on the decision to seek medical treatment.

(2) Yichi Zhang, North Carolina State University
Title: *A Generalized R-Learner for the Heterogeneous Causal Effect Estimation with Continuous Treatments*

Abstract: Estimation of the heterogeneous causal effect is a fundamental topic in causal inference. Recent years have witnessed the growth of the methods for the conditional average treatment effect (CATE) estimation with binary treatments. However, the flexible estimation of CATE with continuous treatments, yet important in practice, is still lacking in the literature. In this paper, we generalize the celebrated R-learner for CATE estimation with binary treatment, to the continuous treatment scenario. A plain extension of the R-learner to the continuous treatment scenario can not identify the CATE due to the ill-posedness. We resolve such identification issues by introducing an additional $\ell_2$ penalty in the R-loss while approximating the target estimand with B-splines. Our new estimator is consistent with the CATE with continuous treatments. Furthermore,

it presents the rate-double robustness, i.e., it has a robust convergence rate even if the two nuisance functions can not be estimated particularly accurately.

(3) Yan Shuo Tan, University of California, Berkeley
Title: *Stable Discovery of Interpretable Subgroups*

Abstract: Building on Yu and Kumbier's PCS framework and for randomized experiments, we introduce a novel methodology for Stable Discovery of Interpretable Subgroups via Calibration (StaDISC), with large heterogeneous treatment effects. StaDISC was developed during our re-analysis of the 1999-2000 VIGOR study, an 8076 patient randomized controlled trial (RCT), that compared the risk of adverse events from a then newly approved drug, Rofecoxib (Vioxx), to that from an older drug Naproxen. Vioxx was found to, on average and in comparison to Naproxen, reduce the risk of gastrointestinal (GI) events but increase the risk of thrombotic cardiovascular (CVT) events. Applying StaDISC, we fit 18 popular conditional average treatment effect (CATE) estimators for both outcomes and use calibration to demonstrate their poor global performance. However, they are locally well-calibrated and stable, enabling the identification of patient groups with larger than (estimated) average treatment effects. In fact, StaDISC discovers three clinically interpretable subgroups each for the GI outcome (totaling 29.4

**Session** 9: ***Statistical Methods for Integrative Data Analysis***
Organizer and Chair: Peter Song, University of Michigan
**Room: Elder Tom Crane Bear, Time: 1:20 PM – 3:00 PM**

(1) Emily Hector, North Carolina State University
Title: *Functional Regression with Wearable Device Data: a New Lens Through Data Partitioning*

Abstract: Modern longitudinal data from wearable devices consist of biological signals at high-frequency time points and offer unparalleled opportunities for discovering new health insights. Distributed statistical methods have emerged as a powerful tool to overcome the computational burden of estimation and inference with these intensively measured outcomes, but methodology for distributed functional regression remains limited. Developing functional regression tools is critical to appropriately modeling and understanding these data. We propose a distributed estimation and inference procedure that efficiently estimates both functional and scalar parameters for intensively measured longitudinal outcomes and overcomes computational difficulties through a scalable divide-and-conquer algorithm. We circumvent traditional basis selection problems by analyzing data in smaller subsets such that the basis functions have a low dimension. To address the challenges of combining estimates from dependent analyses, we propose a statistically efficient one-step estimator that avoids estimation of higher-order moments. We show theoretically and numerically that the proposed estimator is as statistically efficient as a non-distributed approach and more efficient computationally. We demonstrate the practicality of our approach through application of our method to accelerometer data from the NHANES data set.

(2) Lan Luo, The University of Iowa
Title: *Multivariate Online Regression Analysis with Heterogeneous Streaming Data*

Abstract: New data collection and storage technologies have given rise to a new field of streaming data analytics, including real-time statistical methodology for online data analyses. Most existing online learning methods are based on homogeneity assumption such that the sequence of samples are independent and identical. However, inter-data batch correlation and dynamically evolved batch-specific effects are among the key defining features in real-world streaming data such as electronic health records and mobile health data. This talk centers around the state space-mixed models in which the observed data stream is driven by a latent state process that follows a Markov process. In this setting, online maximum likelihood estimation is challenged by high-dimensional integrals and complex covariance structures. In this project, we develop a Kalman filter based real-time regression analysis method that enables to update both point estimates and standard errors of the fixed population average effects while adjusting for dynamic hidden effects. Both theoretical justification and numerical experiments have demonstrated that our proposed online method has similar statistical properties to its offline counterpart but enjoys great computation efficiency. We also apply this method to analyze an electronic health record data example. This is the joint work with Professor Peter X.-K. Song in the Department of Biostatistics at the University of Michigan.

(3) Peisong Han, University of Michigan
Title: *Integrating Summary Information from many External Studies with Heterogeneous Populations*

Abstract: For an internal study of interest, information provided by relevant external studies can be useful to improve the efficiency for parameter estimation in model building, and the external information is oftentimes in summary form. When information is available from possibly many external studies, extra care is needed due to inevitable study population heterogeneity. The information from studies with populations different from the internal study may harm model fitting by introducing estimation bias. We allow the number of external studies that can be considered for possible information integration to increase with the internal sample size, and develop an effective method that integrates only the helpful information for efficiency improvement without introducing bias. Using this method, we study the change of mental health for individuals with bipolar disorder during COVID-19 pandemic, by integrating summary information from relevant existing large-scale studies.

(4) Xiaotian Dai, University of Calgary
Title: *Statistical Framework to Support the Epidemiological Interpretation of SARS-CoV-2 Concentration in Municipal Wastewater*

Abstract: The ribonucleic acid (RNA) of the severe acute respiratory syndrome coronavirus 2 (SARS-Cov-2) is detectable in municipal wastewater as infected individuals can shed the virus in their feces. Viral concentration in wastewater

can inform the severity of the COVID-19 pandemic but observations can be noisy and hamper the epidemiological interpretation. Motivated by a Canadian nationwide wastewater surveillance data set, we aim i) to detect the true trends of viral concentration out of noisy wastewater observations; and ii) to accurately forecast the future trajectory of viral concentrations in wastewater. To achieve these aims, we propose a novel statistical algorithm to model viral concentration in wastewater based on functional data analysis in a Bayesian framework. We apply our model to a nationwide wastewater surveillance program covering 15 sampling sites across Canada. We also assess the performance of our algorithm using simulated data.

**Session** 10: ***Statistical Methods for High Dimensional Data***
Organizer and Chair: Xuekui Zhang, University of Victoria
**Room: MB251, Time: 1:20 PM – 3:00 PM**

(1) Li Xing, University of Saskatchewan
Title: *EPPS: a Novel Ensemble Test to Improve the Power of Genomic Studies*
Abstract: The traditional Genome-Wide Association Studies (GWASs) approach often involves testing associations between a disease outcome and millions of single nucleotide polymorphisms (SNPs). It applies multiple testing adjustments afterwards to control the false positive rate. Due to the curse of dimensionality and the limited sample size, many GWASs face a lack of power issues based on such an approach. We proposed EPPS, a novel ensemble test procedure based on multiple random data splits, which overcomes two drawbacks of the other multiple random data splits approaches. First, EPPS provides 'one' p-value for each SNP by integrating all data splits' results, enabling FDR control using any standard multiple testing adjustment approaches. Second, it automatically selects values of its parameters to optimize study power with a pre-hoc power analysis. We demonstrated that EPPS outperforms the traditional method in simulation and real data application.

(2) Depeng Jiang, University of Manitoba
Title: *Latent Transition Analysis for Multilevel and Multivariate Longitudinal Outcomes*

Abstract: Many intervention programs often have multiple outcome variables, each being reported for multiple time points (e.g., pre and post intervention) where data for participants are organized at more than one level (students nested in schools). Latent class analysis (LCA) is one of the most commonly used cluster technique to identify the meaningful subgroups (latent classes) with multiple indicator variables, while the latent transition analysis (LTA) is a well-known approach for modeling transition of these subgroups over time. Recently, LTA with random intercepts was proposed and has been shown that it can lead to better estimates of the transition probabilities and extract new information from the data. But this has not been extended for LTA with multiple group comparisons in multilevel setting. In this paper, we propose LTA with random intercept variation for program evaluation with multivariate longitudinal data with nested

structure. Between-subject variation is separated from the within-subject latent class transition over time allowing a clear interpretation of the data. The robust standard errors in LTA were considered to take into account the non-independence of observations due to the clustered structure. Data from the Manitoba PAX Study serve as an illustration. PAX is an on-going province-wide classroom based program designed to improve students' mental health in Manitoba. We compare the LTA models with different random intercept variations with the regular LTA. We examine how the mental health states transit from pre- to post-intervention and to follow-up assessment and whether the intervention program affect the transition probabilities. The strengths and limitations of these LTA models are discussed.

(3) Liangliang Wang, Simon Fraser University
Title: *Generalized Bayesian Multidimensional Scaling*
Abstract: Multidimensional scaling is widely used to reconstruct a map with the points' coordinates in a low-dimensional space from the original high-dimensional space while preserving the distances between points. The existing Bayesian approaches for multidimensional scaling are limited to models with Gaussian measurement errors and implementation using Markov chain Monte Carlo algorithms. To overcome these limitations, we first developed a general framework that incorporates non-Gaussian measurement errors and robustness to fit different types of dissimilarities. Then, we proposed an annealed Sequential Monte Carlo algorithm for Bayesian multidimensional scaling inference. This algorithm provides an approximate posterior distribution over the points' coordinates in a low-dimensional space and an unbiased estimator for the marginal likelihood. This study compared the models' performances based on marginal likelihoods, which are byproducts of the annealed Sequential Monte Carlo algorithm. Using synthetic and real data, we demonstrated the effectiveness of the proposed algorithm. We have found that the proposed algorithm outperforms other benchmark algorithms under the same computational budget based on some common metrics used in the literature.

**Coffee Break, MB Central Foyer**

# Parallel Sessions D
# 3:20pm - 5:00pm, Friday, July 8th

**Session** 11: ***Statistics in Biosciences (Sponsored Session)***
Organizer and Chair: Joan Hu, Simon Fraser University
**Room: MB Auditorium, Time: 3:20 PM – 5:00 PM**

(1) Hongzhe Li, University of Pennsylvania
Title: *Estimation and Inference with Proxy Data and Its Genetic Applications*
Abstract: Existing high-dimensional statistical methods are largely established for analyzing individual-level data. In this work, we study estimation and inference for high-dimensional linear models where we only observe "proxy data",

which include the marginal statistics and sample covariance matrix that are computed based on different sets of individuals. We develop a rate optimal method for estimation and inference for the regression coefficient vector and its linear functionals based on the proxy data. Moreover, we show the intrinsic limitations in the proxy-data based inference: the minimax optimal rate for estimation is slower than that in the conventional case where individual data are observed; the power for testing and multiple testing does not go to one as the signal strength goes to infinity. These interesting findings are illustrated through simulation studies and an analysis of a dataset concerning the genetic associations of hindlimb muscle weight in a mouse population.

(2) Yi Xiong, Fred Hutchinson Cancer Research Center
Title: *Statistical Analysis of Recurrent Events from Administrative Databases*
Abstract: Administrative health data in general contain rich information for investigating public health issues. On the other hand, however, many restrictions and regulations apply to their use. The data are usually not in the conventional format since administrative databases are created and maintained to serve non-research purposes and only information for people who seek health services is accessible. In addition, administrative health databases evolve over time and the regulations about their access may change. Motivated by administrative records of emergency department (ED) visits by children and youths in Alberta, we propose novel statistical methods to address two challenges: (i) to evaluate dynamic pattern and impacts with doubly-censored recurrent event data and (ii) to re-calibrate estimators developed based on truncated information by leveraging summary statistics from the population. These methods are justified theoretically and numerically using both simulation and the ED visits data. This is a joint work with Dr. Joan Hu (Simon Fraser University) and Dr. Rhonda Rosychuk (University of Alberta).

(3) Kwun Chuen Gary Chan, University of Washington
Title: *The National Alzheimer's Coordinating Center Data Set and some Associated Statistical Problems*
Abstract: The National Alzheimer's Coordinating Center has coordinated collection of the Uniform Data Set from over 30 Alzheimer's Disease Research Centers (ADRC) in the USA since 2005. Participants with heterogeneous clinical presentations are followed up annually. The neuropsychological test batteries are being updated periodically to reflect the state-of-the-art of Alzheimer's Disease Research. I will present the data structure and several associated statistical problems. A focus will be on a recently developed semi-parametric mixture method for harmonization across different versions of neuropsychological test batteries of the Uniform Data Set.

(4) Zhezhen Jin, Columbia University
Title: *Analysis of Large Data with Subsampling*
Abstract: Analysis of large data is challenging due to its size and computational issues. Subsampling methods and divide-and-conquer procedures are appealing because they ease computational burden. However, it is challenging to preserve

the validity of the resulting estimation and inference. In this talk, we will discuss a perturbation subsampling approach based on independent and identically distributed stochastic weights for the analysis of large data. We justify the method based on optimizing convex objective functions by establishing asymptotic consistency and normality for the resulting estimators. Simulation studies and real data analysis will also be used to illustrate the finite sample performance of the method.

**Session** 12: ***Computations and Theories for Statistical Learning***
Organizer: Dehan Kong, University of Toronto
Chair: Ying Zhou, University of Toronto
**Room: Elder Tom Crane Bear, Time: 3:20 PM – 5:00 PM**

(1) Wenxin Zhou, University of California San Diego
Title: *Robust Estimation and Inference for Joint Quantile and Expected Shortfall Regression*

Abstract: Expected Shortfall (ES), as a financial term, refers to the average return on a risky asset conditional on the return below a certain quantile of its distribution. The latter is also known as the Value-at-Risk (VaR). In their Fundamental Review of the Trading Book (Basel Committee, 2016, 2019), the Basel Committee on Banking Supervision confirmed the replacement of VaR with ES as the standard risk measure in banking and insurance. From a statistical perspective, we consider a linear regression framework that simultaneously models the quantile and the ES of a response variable given a set of covariates. The existing approach is based on minimizing a joint loss function, which is not only discontinuous but also non-convex. This inevitably limits its applicability for analyzing large-scale data. Motivated by the idea of using Neyman-orthogonal scores to reduce sensitivity with respect to nuisance parameters, we propose a computationally efficient two-step procedure and its robust variant for joint quantile and ES regression. Under increasing-dimensional settings, we establish explicit non-asymptotic bounds on estimation and Gaussian approximation errors, which lay the foundation for statistical inference of ES regression. In high-dimensional sparse settings, we study the theoretical properties of regularized two-step ES regression estimator as well as its robust counterpart.

(2) Jessica Gronsbell, University of Toronto
Title: *Towards Efficient Analysis of Electronic Health Records Data*

Abstract: The adoption of electronic health records (EHRs) has generated massive amounts of routinely collected medical data with potential to improve our understanding of healthcare delivery and disease processes. However, the analysis of EHR data remains both practically and methodologically challenging as it is recorded as a byproduct of clinical care and billing, and not for research purposes. For example, outcome information, such as presence of a disease or treatment response, is often missing or poorly annotated in patient records, which brings challenges to statistical learning and inference. In this talk, I will focus on predictive modeling in settings with an extremely limited amount of outcome

information and demonstrate the advantages of semi-supervised learning methods that incorporate large volumes of unlabeled data into model estimation and evaluation.

(3) Xinyi Zhang, University of Toronto
Title: *Fighting Noise with Noise: Causal Inference with Many Candidate Instruments*

Abstract: Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method that constructs pseudo variables to remove irrelevant candidate instruments having spurious correlations with the exposure. Theoretical and synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

(4) Peter Song, University of Michigan
Title: *Distributed Causal Inference in the Presence of Data-Sharing Barriers*

Abstract: Data sharing barriers are paramount challenges arising from multicenter clinical trials where multiple data sources are stored in a distributed fashion at different local study sites. Merging such data sources may become very burdensome when causal inference is of primary interest. We propose a new causal inference framework that avoids the merging of subject-level raw data from multiple sites but needs only the sharing of summary statistics. The proposed collaborative inference enjoys maximal protection of data privacy and minimal sensitivity to unbalanced data distributions across data sources. We show theoretically and numerically that the new distributed causal inference approach has little loss of statistical power compared to the centralized method that requires merging the entire data. We illustrate its performance by simulation experiments and a real-world data example.

**Session** 13: ***Contemporary Advances in Complex Data Analysis***
Organizer: Juxin Liu, University of Saskatchewan
Chair: Guohua Yan, University of New Brunswick
**Room: MB253, Time: 3:20 PM – 5:00 PM**

(1) Guohua Yan, University of New Brunswick
Title: *Binary Logistic Models with Partially Crossed Random Effects*

Abstract: Scientists in many disciplines frequently encounter data with binary outcomes that have cross-classified data structures. For example, in a student

admission study (success or failure), schools and areas could be treated as crossed random effects since not all students from the same school live in the same area and vice versa. It is crucial to incorporate crossed random effects into the model for data with cross-classified structures; otherwise, data analysis results might be misleading. In this talk we propose a binary logistic model with partially crossed random effects. We predict the random effects in the model by the orthodox best linear unbiased predictor (BLUP) approach. The model is robust because it specifies only the first and second moments of the random effects. As an illustration, we analyze an insurance dataset about motor vehicle accidents. This is a joint work with Renjun Ma and Zizhe Zhang.

(2) Longhai Li, University of Saskatchewan
Title: *Residual Diagnostics for Censored Regression*

Abstract: Residuals in normal regression are used to assess a model's goodness-of-fit (GOF) and discover directions for improving the model. However, there is a lack of residuals with a characterized reference distribution for censored regression. In this article, we propose to diagnose censored regression with normalized randomized survival probabilities (RSP). The key idea of RSP is to replace the survival probability (SP) of a censored failure time with a uniform random number between 0 and the SP of the censored time. We prove that RSPs always have the uniform distribution on (0, 1) under the true model with the true generating parameters. Therefore, we can transform RSPs into normally distributed residuals with the normal quantile function. We call such residuals by normalized RSP (NRSP residuals). We conduct simulation studies to investigate the sizes and powers of statistical tests based on NRSP residuals in detecting the incorrect choice of distribution family and nonlinear effect in covariates. Our simulation studies show that, although the GOF tests with NRSP residuals are not as powerful as a traditional GOF test method, a nonlinear test based on NRSP residuals has significantly higher power in detecting nonlinearity. We also compared these model diagnostics methods with a breast-cancer recurrent-free time dataset. The results show that the NRSP residual diagnostics successfully captures a subtle nonlinear relationship in the dataset, which is not detected by the graphical diagnostics with CS residuals and existing GOF tests.

(3) Hui Zhang, Northwestern University
Title: *Unbiased and Robust Analysis of Co-Localization in Super-Resolution Images*

Abstract: Spatial data from high-resolution images abound in many scientific disciplines. For example, single-molecule localization microscopy, such as stochastic optical reconstruction microscopy, provides super-resolution images to help scientists investigate co-localization of proteins and hence their interactions inside cells, which are key events in living cells. However, there are few accurate methods for analyzing co-localization in super-resolution images. The current methods and software are prone to produce false-positive errors and are restricted to only 2-dimensional images. In this paper, we develop a novel statistical method to effectively address the problems of unbiased and robust quantifica-

tion and comparison of protein co-localization for multiple 2- and 3-dimensional image datasets. This method significantly improves the analysis of protein co-localization using super-resolution image data, as shown by its excellent performance in simulation studies and an analysis of light chain 3-lysosomal-associated membrane protein 1 protein co-localization in cell autophagy. Moreover, this method is directly applicable to co-localization analyses in other disciplines, such as diagnostic imaging, epidemiology, environmental science, and ecology.

(4) Fangya Mao, University of Waterloo
Title: *Spatial Dependence Modeling of Susceptibility and Failure Times for Processes under Intermittent Observation*

Abstract: Statistical models which facilitate exploration of spatial dependence can provide important scientific insights into chronic diseases processes affecting several organ systems or body sites. In this work we describe models and methods for studying spatial dependence of joint damage in psoriatic arthritis (PsA). Since a large number of joints may remain unaffected even among individuals with a long disease history, spatial dependence is first modelled in latent joint-specific indicators of susceptibility. Among susceptible joints, a Gaussian copula is adopted for dependence modeling of times to damage. Composite likelihoods are developed for settings where individuals are under intermittent observation and progression times are subject to type K interval censoring. Two-stage estimation procedures help mitigate the computational burden arising when a large number of processes (i.e. joints) are under consideration. Simulation studies confirm that the proposed methods provide valid inference, and an application to the motivating data from the University of Toronto Psoriatic Arthritis Clinic yields important scientific insights which can help physicians distinguish PsA from arthritic conditions with different dependence patterns.

**Session** 14: ***when Statistics Meets Complex Data: New Methods in Statistical Learning and Inference***
Organizer and Chair: Wen Zhou, Colorado State University
**Room: MB252, Time: 3:20 PM – 5:00 PM**

(1) HaiYing Wang, University of Connecticut
Title: *Nonuniform Negative Sampling and Log Odds Correction with Rare Events Data*

Abstract: We investigate the issue of parameter estimation with nonuniform negative sampling for imbalanced data. We first prove that, with imbalanced data, the available information about unknown parameters is only tied to the relatively small number of positive instances, which justifies the usage of negative sampling. However, if the negative instances are subsampled to the same level of the positive cases, there is information loss. To maintain more information, we derive the asymptotic distribution of a general inverse probability weighted (IPW) estimator and obtain the optimal sampling probability that minimizes its variance. To further improve the estimation efficiency over the IPW method, we propose a likelihood-based estimator by correcting log odds for the sampled

data and prove that the improved estimator has the smallest asymptotic variance among a large class of estimators. It is also more robust to pilot misspecification. We validate our approach on simulated data as well as a real click-through rate dataset with more than 0.3 trillion instances, collected over a period of a month. Both theoretical and empirical results demonstrate the effectiveness of our method.

(2) Fang Han, University of Washington
Title: *On Azadkia-Chatterjee's Correlation Coefficients*
Abstract: In recent work, Azadkia and Chatterjee laid out an ingenious approach to defining consistent measures of dependence. Their fully nonparametric approach forms statistics based on ranks and nearest-neighbor graphs. The appealing nonparametric consistency of the resulting dependence measure and the associated empirical correlation coefficient has quickly prompted follow-up work that seeks to study its statistical properties. In this talk, we will summarize recent progress and highlight some open problems along this track.

(3) Zhao Ren, University of Pittsburgh
Title: *Heteroskedastic Sparse PCA in High Dimensions*
Abstract: Principal component analysis (PCA) is one of the most commonly used techniques for dimension reduction and feature extraction. Though it has been well-studied for high-dimensional sparse PCA, little is known when the noise is heteroskedastic, which turns out to be ubiquitous in many scenarios, like biological sequencing data and information network data. We propose an iterative algorithm for sparse PCA in the presence of heteroskedastic noise, which alternatively updates the estimates of the sparse eigenvectors using power method with adaptive thresholdings in one step, and imputes the diagonal values of the sample covariance matrix to reduce the estimation bias due to heteroskedasticity in the other step. Our procedure is computationally fast and provably optimal under the generalized spiked covariance model, assuming the leading eigenvectors are sparse. A comprehensive simulation study demonstrates its robustness and effectiveness under various settings.

(4) Yi Archer Yang, McGill University
Title: *Flexible Regularized Estimating Equations: Some New Perspectives*
Abstract: We make some observations about the equivalences between regularized estimating equations, fixed-point problems and variational inequalities: (a) A regularized estimating equation is equivalent to a fixed-point problem, specified via the proximal operator of the corresponding penalty. (b) A regularized estimating equation is equivalent to a (generalized) variational inequality. Both equivalences extend to any estimating equations with convex penalty functions. To solve large-scale regularized estimating equations, it is worth pursuing computation by exploiting these connections. While fast computational algorithms are less developed for regularized estimating equation, there are many efficient solvers for fixed-point problems and variational inequalities. In this regard, we apply some efficient and scalable solvers which deliver hundred-fold speed improvement. These connections can lead to further research in both computational

and theoretical aspects of the regularized estimating equations.

**Session** 15: ***Recent Advances in Data Science***
Organizer: Dehan Kong/Linglong Kong
Chair: Yichi Zhang, North Carolina State University
**Room: MB251, Time: 3:20 PM – 5:00 PM**

(1) Pang Du, Virginia Tech
Title: *Optimal Rate of Convergence of Multivariate Nonparametric Change Point Detection*

Abstract: Change-point analysis of an unlabeled sample of observations consists in, first, testing whether a change in the distribution occurs within the sample, and second, if a change occurs, estimating the change-point instant after which the distribution of the observations switches from one distribution to another different distribution. Recently, the nonparametric testing are popular to serve the first purpose. However, there is still limited work on studying the convergence rate of the change point since the estimation is usually involved a infinite series of testing statistics. In this paper, we establish a non-asymptotic theory for nonparametric density estimation in a reproducing kernel Hilbert space. Based on the derived non-asymptotic bound, we are able to derive the convergence rate of the proposed change point estimator.

(2) Katarzyna Reluga, University of Toronto
Title: *Post-Selection Inference for Linear Mixed Models*

Abstract: While the post-selection inference has received considerable attention in linear and generalized linear models, it seems to be a neglected topic in the field of mixed models and mixed effect predictions. Therefore we have developed a complete asymptotic theory for post-selection inference within the framework of linear mixed models once the conditional Akaike information criterion was employed as a model selection procedure. Our theory is then used to construct confidence intervals for regression parameters, linear statistics and mixed effects under different scenarios: nested and general model sets as well as sets composed only of misspecified models. The theoretical analysis is accompanied by a simulation study which confirms a good performance of our procedures. Moreover, our simulations reveal a startling robustness of the classical confidence intervals for a mixed parameter. This is in contrast to findings for the fixed parameters and may indicate that, under certain scenarios, random effects would automatically adjust for model selection. We illustrate the utility of our proposed methodology in a study of the body mass index across different subgroups of the US population.

(3) Eardi Lila, University of Washington
Title: *Functional Classification for Manifold Imaging Data*

Abstract: We introduce a novel framework for the classification of imaging data supported on non-linear, and possibly random, manifold domains. The motivating application is the identification of subjects with Alzheimer's disease from

their cortical surface geometry and associated cortical thickness map. The proposed model is based upon a reformulation of the classification problem into a regularized multivariate functional linear regression model. This allows us to adopt a direct approach to the estimation of the most discriminant direction while controlling for its complexity with appropriate differential regularization. We apply the proposed method to a pooled dataset from the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Progression Markers Initiative, and are able to estimate discriminant directions that capture both cortical geometric and thickness predictive features of Alzheimer's Disease.

(4) Yanyuan Ma, Pennsylvania State University
Title: *Network Functional Varying Coefficient Model*

Abstract: We consider functional responses with network dependence observed for each individual at irregular time points. To model both the inter-individual dependence as well as within-individual dynamic correlation, we propose a network functional varying coefficient (NFVC) model. The response of each individual is characterized by a linear combination of responses from its connected nodes and its own exogenous covariates. All the model coefficients are allowed to be time dependent. The NFVC model adds to the richness of both the classical network autoregression model and the functional regression models. To overcome the complexity caused by the network inter-dependence, we devise a special nonparametric least squares type estimator, which is feasible when the responses are observed at irregular time points for different individuals. The estimator takes advantage of the sparsity of the network structure to reduce the computational burden. To further conduct the functional principal component analysis, a novel within-individual covariance function estimation method is proposed and studied. Theoretical properties of our estimators are analyzed, which involve techniques related to empirical processes, nonparametrics, functional data analysis and various concentration inequalities. We analyze a social network data to illustrate the powerfulness of the proposed procedure.

## Poster Session, 5:00 pm – 9:00 pm, MB Central Foyer

(1) Yan Cui, University of Toronto
Title: *Optimal forecast for locally stationary functional time series using double-sieve method*

Abstract: Accurate curve forecasting is of vital importance for policy planning, decision making and resource allocation in many engineering and industrial applications. In this paper we establish a theoretical foundation for the optimal short-term linear prediction of non-stationary functional or curve time series with smoothly time-varying data generating mechanisms. The core of this work is to establish a unified functional auto-regressive approximation result for a general class of non-stationary functional time series. A double sieve expansion method is proposed and theoretically verified for the asymptotic optimal forecasting.

A telecommunication traffic data set is used to illustrate the usefulness of the proposed theory and methodology.

(2) Zehui Wang, Queen's University
Title: *Estimation of Cutpoint for a Continuous Biomarker and Paired Bootstrap Tests for Treatment-Biomarker Interaction Based on a Nonparametric Measure of Treatment Effects with Survival Data*

Abstract: Patients with varying degrees of response to specific treatments may be classified using a biomarker when the treatment-biomarker interaction is significant. The traditional approaches are based on the Cox proportional hazard models, however, the proportional hazard assumption under this model could be violated in practice. For a continuous biomarker, a cutpoint is also required for classification. In this report, instead, based on a non-parametric measure of treatment effect, a procedure is proposed to estimate the optimal cut-point of a continuous biomarker and paired bootstrap tests are introduced to assess the treatment-biomarker interaction effects with respect to censored survival outcomes. The evaluation and comparison of the proposed procedure and tests are conducted through Monte Carlo simulations. The proposed approaches are also applied to a data set from a clinical trial on pancreatic cancer.

(3) Danika Lipman, University of Calgary
Title: *Integrative multi-omic analysis reveals enriched pathways associated with COVID-19 and COVID-19 severity*

Abstract: Severe acute respiratory syndrome coronavirus 2, more commonly known as COVID-19, is a disease which is unique in its unpredictable clinical outcomes. It is of interest to understand the underlying mechanisms of the disease and identify key biomarkers. To understand the molecular signature of the disease we have performed an analysis on a multi-omic dataset containing lipidomic, proteomic, metabolomic, RNAseq, and clinical data for 123 patients experiencing COVID-19 or COVID-19 like symptoms. Of these patients, 99 tested positive for COVID-19 giving a control group of 24. Two integrative analysis approaches are taken: Sparse Integrative Discriminant Analysis (SIDA) and Bayesian Integrative Prediction with Networks analysis (BIPnet). From these methods we are able to determine key molecules in disease status and disease severity, and further identify pathways of molecules which are significant. Some of the molecules we have determined to be significant have been identified in other research, while other molecules are novel to this study. These molecules provide us with information on possible treatments and therapies for those with COVID-19 to have better outcomes. We have also identified enriched pathways in COVID-19 disease, and pathways which are enriched with disease severity. Of these pathways there are many surrounding inflammation and immune response, but also some more interesting pathways which provide insight into possible consequences of the disease. These pathways include the neuroprotective role of THOP1 in Alzheimer's disease, atherosclerosis signaling, and maturity onset diabetes of young (MODY) signalling.

(4) Dila Ram Bhandari, Nepal Commerce Campus, Tribhuvan University

Title: *Statistical Models of Machine Learning and Time Series Analysis for Forecasting Crime Pattern*

Abstract: The 20th century saw the establishment of new states where ethnic, religious, linguistic, caste, communal, tribal and other identities played a role in institution of constitutions and in the legal sphere of criminal and victim justice. Recently, South Asian countries face acute problems of corruption, criminal violence, terrorism, extremism, poverty, environmental degradation, cybercrimes, violations of human rights, terrorism, crime against, individual and collective victimization. Everyday massive number of crimes are steadfast, these frequent crimes have made the lives of common citizens restless. Crimes are one of the major threats to society and also for civilization. Crime is a bone of contention that can create a societal disturbance. The old-style crime solving practices are unable to live up to the requirement of existing crime situations. Crime analysis is one of the most important activities of the majority of the intelligent and law enforcement organizations all over the world. The South Asia region lacks such regional coordination mechanism unlike central Asia of Asia Pacific regions to facilitate criminal intelligence sharing and operational coordination related to organized crime, including illicit drug trafficking and money laundering. Machine learning can play an important role to better understand and analyze the forthcoming trend of violations. Different time-series forecasting models have been used to predict the crime ARIMA model and Exponential Smoothing Methods. These forecasting models are trained to predict future violent crimes. The crime records of 2005-2019 which was collected from Nepal Police headquarter and analysed by R programming.

(5) Kevin Zhang, University of Toronto
Title: *Modelling Cellular Development Trajectory using Unbalanced Optimal Transport*

Abstract: Single cell trajectory analysis attempts to reconstruct the biological developmental processes of cells that undergo changes over time by exploiting temporal correlations among gene expressions. Gene expressions of cells usually change during the cellular development and vary across different types of cells. One major problem is that RNA sequencing kills the cell, which makes it infeasible to track the gene expression of a single cell across different stages. Recent study developed an alternative to model the trajectory of individual cells using tools from optimal transport. In this paper, we instead focus on the groups of cells, namely different cell types, and utilize this tool to study how cell types differentiate throughout the time course. In particular, we develop a change point detection algorithm based on discrete entropy regularized unbalanced optimal transport to detect the time points that the cell types differentiate. We further infer how cell types change to different state based on the transport matrix. We evaluate the proposed method using the single cell RNA data from Mouse Embryonic Fibroblasts.

(6) Mei Li, University of Alberta
Title: *Trustworthy Data-Driven Decision Making via Conditional Stochastic Op-*

*timization*

Abstract: While machine learning systems are used to support intelligent decision-making in diverse research areas, yet concerns regarding model resilience and trustworthiness remain debatable. Typically, a data-driven decision under uncertainty is an operator that maps the random variable into a feasible action. It can always be tasked with optimizing a surrogate optimization constructed independently from the unknown probability measure. However, finding such surrogate optimization can be ambiguous and Pareto inefficient. Furthermore, the existing statistical guarantees significantly rely on the assumption of independent randomness, which is frequently debunked in practice. To address these issues, we propose a novel data-driven framework that minimizes a primary risk measure while enforcing an auxiliary quality measure. Such an approach leverages multiple optimality to ensure reliable performance under dependent data-generating processes. In particular, we examine the properties of Sample Average Approximation (SAA) with correlated samples for Conditional Stochastic Optimization to address the challenges in threat analysis, robust optimization design, and model evaluation. We derive exponentially decay error bounds using a rigorous probabilistic error analysis and also show that SAA retains strong asymptotic consistency. We develop a probabilistic robust data-driven approach for surrogate optimization using chance constraints constructed from statistics that satisfy the Large Deviation Principle. Through experiments on several application domains, we illustrate the advantages of the proposed framework including verifying the theoretical results for SAA with dependent data and demonstrating the reliable performance by chance constraint.

(7) Yuzi Liu, University of Alberta

Title: *Sparse Additive Expectile Regression (SAER) in Reproducing Kernel Hilbert Spaces*

Abstract: Sparse estimation has become a mainstream approach for analyzing high-dimensional data. However, the existing studies focus mainly on estimating the mean function, which may fail to characterize the heteroscedasticity and/or asymmetry in model errors. This motivates us to study the sparse additive model under high dimensionality. This paper investigates the estimation of sparse additive expectile Regression (SAER) in high dimension. We propose a regularized learning approach with a two-fold Lasso-type regularization in a reproducing kernel Hilbert space (RKHS) for SAER, and both slow and sharp oracle inequalities for the excess risk of the proposed estimator has been established. Some simulation studies, as well as a real data example, are carried out to illustrate its finite sample performance.

(8) Bo Pan, University of Alberta

Title: *Sample Average Approximation for Stochastic Optimization with Dependent Data: Performance Guarantees and Tractability*

Abstract: Sample average approximation (SAA), a popular method for tractably solving stochastic optimization problems, enjoys strong asymptotic performance guarantees in settings with independent training samples. However, these guar-

antees are not known to hold generally with dependent samples, such as in online learning with time series data or distributed computing with Markovian training samples. In this paper, we show that SAA remains tractable when the distribution of unknown parameters is only observable through dependent instances and still enjoys asymptotic consistency and finite sample guarantees. Specifically, we provide a rigorous probability error analysis to derive 1 - $\beta$ confidence bounds for the out-of-sample performance of SAA estimators and show that these estimators are asymptotically consistent. We then, using monotone operator theory, study the performance of a class of stochastic first-order algorithms trained on a dependent source of data. We show that approximation error for these algorithms is bounded and concentrates around zero, and establish deviation bounds for iterates when the underlying stochastic process is $\varphi$-mixing. The algorithms presented can be used to handle numerically inconvenient loss functions such as the sum of a smooth and non-smooth function or of non-smooth functions with constraints. To illustrate the usefulness of our results, we present several stochastic versions of popular algorithms such as stochastic proximal gradient descent (S-PGD), stochastic relaxed Peaceman–Rachford splitting algorithms (S-rPRS), and numerical experiment.

(9) Na Zhang, University of Alberta
Title: *Renewable $\ell_1$-regularized linear support vector machine with high-dimensional streaming data*

Abstract: The rapid development of modern data collection techniques brings new challenges to existing classification problems and the storage of such huge entire datasets in memory. It is becoming increasingly urgent to develop online updating approaches. This paper studies the renewable estimation process for linear support vector machine (SVM) in the high-dimensional online estimation setting. The renewable estimation process including online $\ell_1$-regularized and online debiased procedures is feasible for high-dimensional streaming data because the estimators are updated using the current data and historical summary statistics instead of re-accessing the raw entire data. Theoretically, we establish the convergence rates of the proposed online estimators under mild conditions. Numerical studies demonstrate the effectiveness of the proposed procedures.

(10) Enze Shi, University of Alberta
Title: *An adaptive model checking test for functional linear model*

Abstract: Numerous studies have been devoted to the estimation and inference problems for functional linear models (FLM). However, few works focus on model checking problem that ensures the reliability of results. Limited tests in this area do not have tractable null distributions or asymptotic analysis under alternatives. Also, the functional predictor is usually assumed to be fully observed, which is impractical. To address these problems, we propose an adaptive model checking test for FLM. It combines regular moment-based and conditional moment-based tests, and achieves model adaptivity via the dimension of a residual-based subspace. The advantages of our test are manifold. First, it has a tractable chi-squared null distribution and higher powers under the alternatives

than its components. Second, asymptotic properties under different underlying models are developed, including the unvisited local alternatives. Third, the test statistic is constructed upon finite grid points, which incorporates the discrete nature of collected data. We develop the desirable relationship between sample size and number of grid points to maintain the asymptotic properties. Besides, we provide a data-driven approach to estimate the dimension leading to model adaptivity, which is promising in sufficient dimension reduction. We conduct comprehensive numerical experiments to demonstrate the advantages the test inherits from its two simple components.

(11) Lei Ding, University of Alberta
Title: *Word Embeddings via Causal Inference: Gender Bias Reducing and Semantic Information Preserving*

Abstract: 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.

(12) Ke Sun, University of Alberta
Title: *Exploring the Training Robustness of Distributional Reinforcement Learning against Noisy State Observations*

Abstract: In real scenarios, state observations that an agent observes may contain measurement errors or adversarial noises, misleading the agent to take suboptimal actions or even collapse while training. In this paper, we study the training robustness of distributional Reinforcement Learning (RL), a class of state-of-the-art methods that estimate the whole distribution, as opposed to only the expectation, of the total return. Firstly, we validate the contraction of both expectation-based and distributional Bellman operators in the State-Noisy Markov Decision Process (SN-MDP), a typical tabular case that incorporates both random and adversarial state observation noises. Beyond SN-MDP, we then analyze the vulnerability of least squared loss in expectation-based RL with either linear or nonlinear function approximation. By contrast, we theoretically characterize the bounded gradient norm of distributional RL loss based on the categorical parameterization. The resulting stable gradients while the optimization in distributional RL accounts for its better training robustness against state

observation noises. Finally, extensive experiments on the suite of games verified the convergence of both expectation-based and distributional RL in the SN-MDP setting under different strengths of state observation noises. More importantly, in noisy settings beyond SN-MDP, distributional RL is less vulnerable against noisy state observations compared with its expectation-based counterpart.

**Reception, 6:00 pm – 9:00 pm, MB Central Foyer**

# Plenary Talk II
# 8:30am - 9:30am, Saturday, July 9th

**Session** 16: ***Keynote Speech 2***
Organizer: Yingwei Peng, Queen's University
Chair: Linglong Kong, University of Alberta
**Room: MB Auditorium, Time: 8:30 AM – 9:30 AM**

(1) Jianqing Fan, Princeton University
Title: *The Efficacy of Pessimism in Asynchronous Q-Learning*
Abstract: This paper is concerned with the asynchronous form of Q-learning, which applies a stochastic approximation scheme to Markovian data samples. Motivated by the recent advances in offline reinforcement learning, we develop an algorithmic framework that incorporates the principle of pessimism into asynchronous Q-learning, which penalizes infrequently-visited state-action pairs based on suitable lower confidence bounds (LCBs). This framework leads to, among other things, improved sample efficiency and enhanced adaptivity in the presence of near-expert data. Our approach permits the observed data in some important scenarios to cover only partial state-action space, which is in stark contrast to prior theory that requires uniform coverage of all state-action pairs. When coupled with the idea of variance reduction, asynchronous Q-learning with LCB penalization achieves near-optimal sample complexity, provided that the target accuracy level is small enough. In comparison, prior works were suboptimal in terms of the dependency on the effective horizon even when i.i.d. sampling is permitted. Our results deliver the first theoretical support for the use of pessimism principle in the presence of Markovian non-i.i.d. data. (Joint with Yuling Yan, Gen Li, and Yuxin Chen)

**Coffee Break, MB Central Foyer**

# Parallel Sessions F
# 9:50am - 11:30pm, Saturday, July 9th

**Session** 17: ***Challenges in Modern Data Analysis and Reproducibility***
Organizer: Bei Jiang, University of Alberta
Chair: Jinhan Xie, University of Alberta
**Room: MB Auditorium, Time: 9:50 AM – 11:30 AM**

(1) Yao Luo, University of Toronto
Title: *Penalized Sieve Estimation of Structural Models*
Abstract: Estimating structural models is an essential tool for economists. However, existing methods are often inefficient either computationally or statistically, depending on how equilibrium conditions are imposed. We propose a class of penalized sieve estimators that are consistent, asymptotic normal, and asymptotically efficient. Instead of solving the model repeatedly, we approximate the solution with a linear combination of basis functions and impose equilibrium conditions as a penalty in searching for the best fitting coefficients. We apply our method to an entry game between Walmart and Kmart.

(2) Yeying Zhu, University of Waterloo
Title: *Causal Mediation Analysis with Multiple Mediators*
Abstract: Causal mediation analysis has become popular in recent years. The goal of mediation analyses is to learn the direct effects of exposure on outcome as well as mediated effects on the pathway from exposure to outcome. Very often, the indirect pathway is carried out by more than one mediators. In this talk, we discuss joint modelling approaches for estimating direct and indirect effects using flexible statistical models that account for correlations among the mediators. Valid inference regarding the estimated direct and indirect effects will be discussed. Extensive simulation studies are conducted and the proposed methods are applied to an epigenetic study in which the goal is to understand how DNA methylation mediates the effect of childhood trauma on regulation of human stress reactivity.

(3) Xiaodong Yan, Shandong University
Title: *Bandit Inference for Small Group Treatment Effect*
Abstract: Enhancing the power of treatment effect testing statistics attracts essential attention among sorts of discipline areas, especially in small group. Motivated by two armed bandit process, this article proposes a strategic sampling procedure to construct a treatment effect testing statistics by combing parts of the law of large numbers and central limit theorem, and employs nonlinear limit theory to study its asymptotic behaviour referred to strategic central limit theorem (strategic CLT) which is the original theory we propose. We also provide a common strategic sampling-based bootstrap to recover the limit distribution of the developed statistics, making its use possible on observational dataset and scalable for other hypothesis testings. The theoretical results achieve the explicit density function of limit distribution, known as bandit distribution. Simulation studies pose supportive evidence that the proposed spike statistics performs well with finite samples and especially shows powerful behavior with small size of the sample. A real data example is provided for illustration.

(4) Radu Craiu, University of Toronto
Title: *General Behaviour of P-Values Under the Null and Alternative*
Abstract: Hypothesis testing results often rely on simple, yet important assumptions about the behaviour of the distribution of p-values under the null and

alternative. We examine tests for one dimensional parameters of interest that converge to a normal distribution, possibly in the presence of many nuisance parameters, and characterize the distribution of the p-values using techniques from the higher order asymptotics literature. We show that commonly held beliefs regarding the distribution of p-values are misleading when the variance or location of the test statistic is not well-calibrated or when the higher order cumulants of the test statistic are not negligible. We further examine the impact of having these misleading p-values on reproducibility of scientific studies, with some examples focused on GWAS studies. Corrected tests are proposed and are shown to perform better than their traditional counterparts in various settings. This is joint work with Yanbo Tang and Lei Sun.

**Session** 18: ***Statistical Considerations in Complex Biomedical Data Analysis***
Organizer and Chair: Weining Shen, University of California, Irvine
**Room: MB252, Time: 9:50 AM – 11:30 AM**

(1) Jin Zhou, University of California, Los Angeles
Title: *GWAS of Longitudinal Trajectories at Biobank Scale*
Abstract: Biobanks linked to massive, longitudinal electronic health record (EHR) data make numerous new genetic research questions feasible. One among these is the study of biomarker trajectories. For example, high blood pressure measurements over visits strongly predict stroke onset, and consistently high fasting glucose and Hb1Ac levels define diabetes. Recent research reveals that not only the mean level of biomarker trajectories but also their fluctuations, or within-subject (WS) variability, are risk factors for many diseases. Glycemic vari- ation, for instance, is recently considered an important clinical metric in diabetes management. It is crucial to identify the genetic factors that shift the mean or alter the WS variability of a biomarker trajectory. Compared to traditional cross-sectional studies, trajectory anal- ysis utilizes more data points and captures a complete picture of the impact of time-varying factors, including medication history and lifestyle. Currently, there are no efficient tools for genome-wide association studies (GWASs) of biomarker trajectories at the biobank scale, even for just mean effects. We propose TrajGWAS, a linear mixed effect model-based method for testing genetic effects that shift the mean or alter the WS variability of a biomarker trajectory. It is scalable to biobank data with 100,000 to 1,000,000 individuals and many longitudinal measurements and robust to distributional assumptions. Simulation studies corroborate that TrajGWAS controls the type I error rate and is powerful. Analysis of eleven biomarkers measured longitudinally and extracted from UK Biobank primary care data for more than 150,000 participants with 1,800,000 observations reveals loci that significantly alter the mean or WS variability.

(2) Zhaoxia Yu, University of California Irvine
Title: *Penalized Hypothesis Testing in High-Dimensional Settings*
Abstract: High-dimensionality is ubiquitous in various scientific fields such as imaging genetics, where a deluge of functional and structural data on brain-

relevant genetic polymorphisms are investigated. It is crucial to identify which genetic variations are consequential in identifying neurological features of brain connectivity compared to merely random noise. Statistical inference in high-dimensional settings poses multiple challenges involving analytical and computational complexity. A widely implemented strategy in addressing inference goals is penalized inference. In particular, the role of the ridge penalty in high-dimensional prediction and estimation has been actively studied in the past several years. This study focuses on ridge-penalized tests in high-dimensional hypothesis testing problems by proposing and examining a class of methods for choosing the optimal ridge penalty. We present our findings on strategies to improve the statistical power of ridge-penalized tests and what determines the optimal ridge penalty for hypothesis testing. The application of our work to an imaging genetics study and biological research will be presented.

(3) LAN XUE, Oregon State University
Title: *Local Signal Detection on Irregular Domains with Spatially Varying Coefficient Model*

Abstract: In spatial data analysis, it is essential to understand and quantify spatial or temporal heterogeneity. In this paper, we focus on a spatially varying coefficient model, in which spatial heterogeneity is accommodated by allowing the regression coefficients to vary in a given spatial domain. We propose a model selection method for spatially varying coefficient models using penalized bivariate splines. It uses bivariate splines defined on triangulation to approximate nonparametric varying coefficient functions and minimizes the sum of squared errors with local penalty on L2 norms of spline coefficients for each triangle. Our method partitions the region of interest using triangulation and provides efficient approximation of irregular domains. In addition, we propose an efficient algorithm to obtain the proposed estimator using the local quadratic approximation. We also establish the consistency of estimated nonparametric coefficient functions and the estimated null regions. Moreover, we develop model confidence regions as the inference tool to quantify the uncertainty of the estimated null regions. The numerical performance of the proposed method is evaluated in both simulation case and real data analysis.

(4) Qingxia Chen, VUMC
Title: *Estimation of Treatment Effects and Model Diagnostics with Two-Way Time-Varying Treatment Switching: An Application to a Head and Neck Study*

Abstract: Treatment switching frequently occurs in clinical trials due to ethical reasons. Intent-to-treat analysis without adjusting for switching yields biased and inefficient estimates of the treatment effects. In this paper, we propose a class of semiparametric semicompeting risks transition survival models to accommodate two-way time-varying switching. Theoretical properties of the proposed method are examined. An efficient expectation–maximization algorithm is derived to obtain maximum likelihood estimates and model diagnostic tools. Existing software is used to implement the algorithm. Simulation studies are conducted to demonstrate the validity of the model. The proposed method is

further applied to data from a clinical trial with patients having recurrent or metastatic squamous-cell carcinoma of head and neck.

**Session** 19: ***Modern Statistical Machine Learning in Medicine***
Organizer: Xiaofeng Wang, Cleveland Clinic
Chair: Lingsong Zhang, Purdue University
**Room: MB251, Time: 9:50 AM – 11:30 AM**

(1) linglong Kong, University of Alberta
Title: *Gaussian Copula Function-on-Scalar Regression in Reproducing Kernel Hilbert Space*

Abstract: To relax the linear assumption in function-on-scalar regression, we borrow the strength of copula and propose a novel Gaussian copula function-on-scalar regression. Our model is more flexible to characterize the dynamic relationship between functional response and scalar predictors. Estimation and prediction are fully investigated. We develop a closed form for the estimator of coefficient functions in a reproducing kernel Hilbert space without the knowledge of marginal transformations. Valid, distirbution-free, finite-sample prediction bands are constructed via conformal prediction. Theoretically, we establish the optimal convergence rate on the estimation of coefficient functions and show that our proposed estimator is rate-optimal under fixed and random designs. The finite-sample performance is investigated through simulations and illustrated in real data analysis.

(2) Tingting Zhang, University of Pittsburgh
Title: *A Variational Bayesian Approach to Identifying Whole-Brain Directed Networks with fMRI Data*

Abstract: The brain is a high-dimensional directed network system as it consists of many regions as network nodes that exert influence on each other. The directed influence exerted by one region on another is referred to as directed connectivity. We aim to reveal whole-brain directed networks based on resting-state functional magnetic resonance imaging (fMRI) data of many subjects. However, it is both statistically and computationally challenging to produce scientifically meaningful estimates of whole-brain directed networks. To address the statistical modeling challenge, we assume modular brain networks, which reflect functional specialization and functional integration of the brain. We address the computational challenge by developing a variational Bayesian method to estimate the new model. We apply our method to resting-state fMRI data of many subjects and identify modules and directed connections in whole-brain directed networks. The identified modules are accordant with functional brain systems specialized for different functions. We also detect directed connections between functionally specialized modules, which is not attainable by existing network methods based on functional connectivity. In summary, this paper presents a new computationally efficient and flexible method for directed network studies of the brain as well as new scientific findings regarding the functional organization of the human brain.

(3) Lingsong Zhang, Purdue University
Title: *Generative Models for Diabetic Retinopathy Data*
Abstract: In this talk, analysis of diabetic retinopathy data will be presented. A two stage approach will be discussed. First stage, a novel generative model will be used to simulate the vessel structure for diabetic retinopathy. One the second stage, another different generative model will be proposed to generate the details and other information based on different stage of the diabetes. Extensive simulation will be used and a new measure of the perforamnce will be discussed in this talk as well. This is a joint work with Mingxuan Zhang.

(4) Xiaofeng Wang, Cleveland Clinic
Title: *High-Dimensional Variable Selection and Estimation in Functional Cox Models*
Abstract: In critical care medicine, high-dimensional functional variables and scalar variables are often simultaneously collected. This paper is to study variable selection and estimation for functional Cox models with right-censored data in the presence of both high-dimensional functional and scalar covariates. We investigate and evaluate three group penalized methods for functional Cox models: the group lasso and two nonconvex penalization methods - group SCAD and group MCP. To stabilize the estimation and account for the variation induced by variable selection, we adapt the "splitting and smoothing" algorithm (Fei and Li, 2021) to the estimation of functional Cox models. The sample is shuffle-split into two parts; then, the variable selection is applied using one part, and a partial functional cox regression is conducted using the other part. Averaging the estimates over multiple shuffle splits, we obtain the smoothed estimates. These smoothed estimators are consistent and numerically stable. Comprehensive simulation studies demonstrate the numerical properties of model implementations with the group lasso, SCAD, and MCP regularization terms. Finally, we apply our methods to the data of critically ill patients with Covid-19 at Cleveland Clinic.

**Session** 20: ***Recent Advances in Statistical Genetics***
Organizer: Dehan Kong, University of Toronto
Chair: Kaiqiong Zhao, University of Alberta
**Room: Elder Tom Crane Bear, Time: 9:50 AM – 11:30 AM**

(1) Lei Sun, University of Toronto
Title: *One Step Forward Two Steps Back: Recent Advances and New Challenges in the Analysis of the X Chromosome*
Abstract: The inclusion of the X chromosome (Xchr) in genome-wide association studies is known to be difficult due to multiple analytical challenges, particularly the uncertainty of X-inactivation, where one of the two Xchrs in a female may be randomly or preferentially selected to have no effect (i.e. dosage compensation), and there is also the possibility of no X-inactivation (i.e. X-inactivation escape). To date, only 0.5

In this talk, I will first present an Xchr association method that is robust to X-inactivation uncertainty and easy-to-implement, and compare it with other methods that have focused on X-inactivation. I will then present evidence for the previously under-appreciated phenomenon of sex differences in minor allele frequency (sdMAF), from a recent analysis of the 1000 Genomes Project data and the high coverage whole genome sequence data from gnomAD V 3.1.2. sdMAF may affect the validity and power of the existing X-inactivation-focused association methods, as well as our current (lack of) understanding of Hardy-Weinberg equilibrium and linkage disequilibrium on the Xchr.

(2) Yue Niu, University of Arizona
Title: *Inference for Gaussian Multiple Change-Point Model via Bayesian Information Criterion*
Abstract: For a change-point model with a piecewise constant mean structure and additive Gaussian noises, a fundamental inference problem is to determine the existence of change points. Early works usually assume that there is at most one change point. Many recent works can handle multiple changes, however, mainly focus on identifying individual change points. In particular, it is not clear the weakest condition to guarantee the existence of an asymptotically powerful test. In this talk, we answer this question via a Bayesian information criterion approach.

(3) Michael Wu, Fred Hutchinson Cancer Center
Title: *Kernel-Based Genetic Association Analysis for Microbiome Phenotypes Identifies Host Genetic Drivers of Beta-Diversity*
Abstract: Understanding human genetic influences on the gut microbiota helps elucidate the mechanisms by which genetics affects health outcomes. We propose a novel approach, the covariate-adjusted kernel RV (KRV) framework, to map genetic variants associated with microbiome beta-diversity, which focuses on overall shifts in the microbiota. The proposed KRV framework improves statistical power by capturing intrinsic structure within the genetic and microbiome data while reducing the multiple-testing burden. We apply the covariate-adjusted KRV test to the Hispanic Community Health Study/Study of Latinos in a genome-wide association analysis (first gene-level, then variant-level) for microbiome beta-diversity. We have identified an immunity-related gene, IL23R, reported in previous association studies and discovered 3 other novel genes, 2 of which are involved in immune functions or autoimmune disorders. Our findings highlight the value of the KRV as a powerful microbiome GWAS approach and support an important role of immunity-related genes in shaping the gut microbiome composition.

(4) Kai Wang, University of Iowa
Title: *Two Sample Two Stage Least Squares Mendelian Randomization using Summary Statistics from Heterogeneous Samples*
Abstract: Mendelian randomization is gaining popularity for studying the causal effect of an exposure on an outcome. The two stage least squares (TSLS) regression is a useful technique and is typically approximated by the inverse variance

weighting (IVW) method from meta analysis. I will introduce some of my recent work: 1) The TSLS estimate can be computed from GWAS summary statistics without using the IVW approximation. 2) The heterogeneity of two samples can be checked graphically. 3) In the presence of two heterogeneity samples, the MR-Egger regression can be used after adopting appropriate weights. These methods will be demonstrated using empirical examples.

**Session** 21: ***Complex and Mass Data Learning***
Organizer: Linglong Kong, University of Alberta
Chair: Dengdeng Yu, University of Texas at Arlington
**Room: MB253, Time: 9:50 AM – 11:30 AM**

(1) Lingzhu Li, University of Alberta
Title: *An Adaptive Model Checking Test for Functional Linear Model*
Abstract: We propose an adaptive model checking test for FLM. It combines regular moment-based and conditional moment-based tests and achieves model adaptivity via the dimension of a residual-based subspace. The advantages of our test are manifold. First, it has a tractable chi-squared null distribution and higher powers under the alternatives than its components. Second, asymptotic properties under different underlying models are developed, including the unvisited local alternatives. Third, the test statistic is constructed upon finite grid points, which incorporates the discrete nature of collected data. We develop the desirable relationship between sample size and the number of grid points to maintain the asymptotic properties. Besides, we provide a data-driven approach to estimate the dimension leading to model adaptivity, which is promising in sufficient dimension reduction. We conduct comprehensive numerical experiments to demonstrate the advantages the test inherits from its two simple components.

(2) Will Wei Sun, Purdue University
Title: *Stochastic Low-Rank Tensor Bandits for Multi-Dimensional Online Decision Making*
Abstract: Multi-dimensional online decision making plays a crucial role in many real applications such as online recommendation and digital marketing. In these problems, a decision at each time is a combination of choices from different types of entities. To solve it, we introduce stochastic low-rank tensor bandits, a class of bandits whose mean rewards can be represented as a low-rank tensor. We consider two settings, tensor bandits without context and tensor bandits with context. In the first setting, the platform aims to find the optimal decision with the highest expected reward, a.k.a, the largest entry of true reward tensor. In the second setting, some modes of the tensor are contexts and the rest modes are decisions, and the goal is to find the optimal decision given the contextual information. We propose two learning algorithms tensor elimination and tensor epoch-greedy for tensor bandits without context, and derive finite-time regret bounds for them. Comparing with existing competitive methods, tensor elimination has the best overall regret bound and tensor epoch-greedy has a sharper dependency on dimensions of the reward tensor. Furthermore, we de-

velop a practically effective Bayesian algorithm called tensor ensemble sampling for tensor bandits with context. Numerical experiments back up our theoretical findings and show that our algorithms outperform various state-of-the-art approaches that ignore the tensor low-rank structure. In an online advertising application with contextual information, our tensor ensemble sampling reduces the cumulative regret by 75

(3) Chad He, Fred Hutchinson Cancer Research Center
Title: *Subtype Analysis with Somatic Mutations*
Abstract: Understanding the association between cancers subtypes and genetic variations is fundamental to the development of targeted therapies for patients. Somatic mutation plays important roles in tumor development and has emerged as a new type of genetic variations for studying the association with cancer subtypes. We propose an approach, SASOM, for the association analysis of cancer subtypes with somatic mutations. Our approach tests the association between a set of somatic mutations (from a genetic pathway) and subtypes, while incorporating functional information of the mutations into the analysis. In a real data application, we examine the somatic mutations from a cutaneous melanoma dataset, and identify a genetic pathway that is associated with immune-related subtypes.

(4) Farouk Nathoo, Mathematics and Statistics, UVIC
Title: *Ant Colony System Optimization for Spatiotemporal Modelling of Combined EEG and MEG Data*
Abstract: Electroencephalography/Magnetoencephalography (EEG/MEG) source localization involves the estimation of neural activity inside the brain volume that underlies the EEG/MEG measures observed at the sensor array. In this paper, we consider a Bayesian finite spatial mixture model for source reconstruction and implement Ant Colony System (ACS) optimization coupled with Iterated Conditional Modes (ICM) for computing estimates of the neural source activity. Our approach is evaluated using simulation studies and a real data application in which we implement a nonparametric bootstrap for interval estimation. We demonstrate improved performance of the ACS-ICM algorithm as compared to existing methodology for the same spatiotemporal model.

## Lunch, 11:30 AM - 1:20 PM, Vistas Dining Room

# Parallel Sessions G
# 1:20pm - 3:00pm, Saturday, July 9th

**Session** 22: ***Statistical Learning in Modern Data Analysis***
Organizer: Linglong Kong, University of Alberta
Chair: Xiaodong Yan, Shandong University
**Room: MB Auditorium, Time: 1:20 PM – 3:00 PM**

(1) Hua Zhou, University of California, Los Angeles
Title: *A Robust Joint Model of Longitudinal Trajectories and Time-to-Event*

*Data at Biobank Scale*

Abstract: Motivated by the analysis of massive electronic health record (EHR) and wearable device data in modern biobanks, we propose a robust and scalable M-estimator, termed the joint model robust estimator (JMRE), for estimating the accelerated failure time (AFT) model for a right-censored event time jointly with a linear mixed model (LMM) for the longitudinal biomarker trajectory. As a semiparametric estimator, JMRE is robust to distribution misspecification in both AFT and LMM models; scalable to biobank data with $10^5 \sim 10^8$ individuals, intensive longitudinal measurements, and a large number of random effects; able to model the time-varying effects on both mean and within-subject variance of the longitudinal biomarker simultaneously; and easily extensible to data with multiple longitudinal biomarkers.

(2) Xiao Wang, Purdue University
Title: *Efficient Multimodal Sampling via Tempered Distribution Flow*

Abstract: Sampling from high-dimensional distributions is a fundamental problem in statistical research and practice, and has become a central task in Bayesian computing, Monte Carlo simulation, and energy-based models. However, great challenges emerge when the target density function is unnormalized and contains multiple modes that are isolated with each other. We tackle this difficulty by fitting an invertible transformation mapping applied to the target distribution, such that the original distribution is warped into a new one that is much easier to sample from. The transformation mapping is constructed based on the normalizing flow model in deep learning. To address the multi-modality issue, our method adaptively learns a sequence of tempered distributions, which we term as a tempered distribution flow, to progressively approach the original distribution. Various experiments demonstrate the superior performance of this novel sampler compared to traditional methods. This is a joint work with Yixuan Qiu.

(3) Hongtu Zhu, The University of North Carolina at Chapel Hill
Title: *Biobank-Scale Multi-Organ Imaging Genetics and Beyond*

Abstract: Challenges in Biobank-scale Brain Imaging Genetics Abstract: Recently the UK Biobank study has conducted brain magnetic resonance imaging (MRI) scans of over 40,000 participants. In addition, publicly available imaging genetic datasets also emerge from several other independent studies. We collected massive individual-level MRI data from different data resources, harmonized image processing procedures, and conducted the largest genetic studies so far for various neuroimaging traits from different structural and functional modalities. In this talk, we showcase novel clinical findings from our analyses, such as the shared genetic influences among brain structures, functions, and the genetic overlaps with a wide spectrum of clinical outcomes. We also discuss challenges we have faced when analyzing these biobank-scale datasets and highlight opportunities for future research. This presentation is based on a series of works with members in the BIG-S2 lab of the University of North Carolina at Chapel Hill. Our results can be easily browsed through the Brain Imaging Genetics Knowledge Portal (BIGKP) (`https://bigkp.org/`).

(4) Ruoqing Zhu, University of Illinois Urbana-Champaign
Title: *Proximal Temporal Consistent Learning for Estimating Infinite Horizon Dynamic Treatment Regimes*

Abstract: Recent advances in mobile health (mHealth) technology provide an effective way to monitor individuals' health statuses and deliver just-in-time personalized interventions. However, the practical use of mHealth technology raises unique challenges to existing methodologies for learning an optimal dynamic treatment regime. Many mHealth applications involve decision-making with large numbers of intervention options and under an infinite time horizon setting where the number of decision stages diverges to infinity. In addition, temporary medication shortages may cause optimal treatments to be unavailable, while it is unclear what alternatives can be used. To address these challenges, we propose a Proximal Temporal consistency Learning (pT-Learning) framework to estimate an optimal regime that is adaptively adjusted between deterministic and stochastic sparse policy models. The resulting minimax estimator avoids the double sampling issue in the existing algorithms. It can be further simplified and can easily incorporate off-policy data without mismatched distribution corrections. We study the theoretical properties of the sparse policy and establish finite-sample bounds on the excess risk and performance error. The proposed method is implemented by our proximalDTR package and is evaluated through extensive simulation studies and the OhioT1DM mHealth dataset.

**Session** 23: ***Recent Development of Statistical Methods for the Analysis of High-Dimensional Data***
Organizer and Chair: Longhai Li, University of Saskatchewan
**Room: MB253, Time: 1:20 PM – 3:00 PM**

(1) Wei Xu, University of Toronto
Title: *Machine-Learning Methodology Development on Disease Prediction using Microbiome Sequence Data*

Abstract: Researchers find that human microbiome is dynamic in nature, attributing to the presence of interactions among microbes, host, and the environment. Multiple research studies have shown that the microbiome is related to disease risk and outcomes. Besides that, microbiome can change over time, by infections or due to medical interventions such as antibiotics. In this presentation, I will introduce some machine learning model development for disease prediction using microbiome sequence data, and how to utilize longitudinal data in disease prediction using repeated microbiome measures. I will show how advanced neural networks such as stratified Convolutional Neural Networks (CNNs) and Long Short-Term Memory networks (LSTMs) can be used for feature extraction and temporal dependency analysis in longitudinal microbiome data. I will also discuss about the challenges and future directions in this research area, along with the performance comparison with other machine learning models.

(2) Haihan Xie, University of Alberta
Title: *Differentially Private Regularized Stochastic Convex Optimization with*

*Heavy-Tailed Data*

Abstract: Existing privacy guarantees for convex optimization algorithms currently do not apply to heavy-tailed data with regularized estimation. This is a notable gap in the differential privacy (DP) literature, given the broad prevalence of non-Gaussian data and penalized optimization problems. In this work, we propose three $(\epsilon, \delta)$-DP methods for regularized convex optimization and derive bounds on their population excess risks in a framework that accommodates heavy-tailed data with fewer assumptions (relative to previous works). This work is the first to address DP in generic convex optimization problems with a nonsmooth regularization term. Two of our methods augment a basic $(\epsilon, \delta)$-DP algorithm with robust procedures for privately estimating minibatch gradients. Our numerical analyses highlight the performance of our methods relative to data dimensionality, batch size, and privacy budget, and suggest settings where each approach is favourable.

(3) Asif Neloy, University of Manitoba
Title: *Auto-Encoders for Anomaly Detection: Efficiency and Trade-Offs*

Abstract: Anomaly detection (AD) is an important research area with diverse application domains: cybersecurity, finance, medical sciences, risk management, etc. AD presents many challenges to the data analyst: the data is generally high-dimensional; anomalies can be quite heterogeneous; they are rare; and they typically arise from unknown data-generating mechanisms. Deep learning and generative modelling have provided a promising approach to address these challenges. By learning the distribution of normal data, anomaly scores can be developed by comparing observations with how well they can be reconstructed by the model. In particular, many auto-encoders have been proposed that seem to perform well on given datasets. However, it can be difficult to compare and contrast these different architectures. In this talk, we present our review of 11 auto-encoder architectures for anomaly detection, divided into three categories: classical models, variational models, and regularized models. We use the MNIST and Fashion-MNIST datasets to assess the following characteristics: reconstruction ability, sample generation, latent space visualization, and anomaly detection accuracy. During our experimentation, we also carefully observed the scope of reproducibility with different training parameters. Using these results, we discuss the efficiency and trade-offs of each architecture.

(4) Duncan Fong, Pennsylvania State University
Title: *A Generalized Ordinal Finite Mixture Regression Model for Market Segmentation*

Abstract: Model-based market segmentation analyses often involve an ordinal dependent variable as ordinal responses are frequently collected in marketing research. In the Bayesian segmentation literature, there are models for an interval- or ratio-scaled dependent variable but there is not any general model for an ordinal dependent variable. In this manuscript, the authors propose a new Bayesian procedure to simultaneously perform segmentation and ordinal regression with variable selection within each derived segment. The procedure is robust to out-

liers and it also provides an option to include concomitant variables that allows the simultaneous profiling of the derived segments. The authors demonstrate that the practice of treating ordinal responses as interval- or ratio-scales to apply existing Bayesian segmentation procedures can lead to very misleading results and conclusions. Through simulation studies, the authors show that the proposed procedure outperforms several benchmark Bayesian segmentation models in parameter recovery, segment retention, and segment membership prediction for such data. Finally, they provide a commercial business customer satisfaction empirical application to illustrate the usefulness of the proposed model.

**Session** 24: ***Advances in Statistical Modeling and Computing for Complex Data***
Organizer: Weixin Yao, University of California, Riverside
Chair: Esra Kurum, University of California, Riverside
**Room: MB252, Time: 1:20 PM – 3:00 PM**

(1) Xinping Cui, University of California, Riverside
Title: *Learning Interactions in Reaction Diffusion Equation with Neural Network*
Abstract: Nonlinear Reaction-diffusion equations are mathematical models that are extensively used in scientific fields. The question of recovering these PDEs based on experimental data is very important but is still widely open in general situations. Recently, significant advances have been made towards this question by taking advantage of methods from deep learning. In literature, most results are concerning the case in which the nonlinear reaction terms are polynomials of the unknown functions. In this talk, we study more complicated equations where the interactions between species exhibit saturation effect (such as logistic function). We propose to learn them as fractions of polynomials. Such equations often appear in activator-inhibitor systems, such as the Gierer-Meinhardt system and the Thomas system. By combining a the modified PDE-Net method and some sparsity analysis, we manage to discover the hidden terms, in particular the fractional terms, as well as their coefficients in these equations according to the simulated data.

(2) Esra Kurum, University of California, Riverside
Title: *A Bayesian Multilevel Time-Varying Framework for Joint Modeling of Hospitalization and Survival in Patients on Dialysis*
Abstract: Over 782,000 individuals in the U.S. have end-stage kidney disease with about 72

(3) Luca Bagnato, Università Cattolica Del Sacro Cuore
Title: *Dimension-Wise Scaled Normal Mixtures*
Abstract: In this work we propose the family of dimension-wise scaled normal mixtures (DSNMs) to model the joint distribution of a d-variate random variable with real-valued components. Each member of the family generalizes the multivariate normal (MN) distribution in two directions. Firstly, the DSNM has a more general type of symmetry with respect to the elliptical symmetry of the MN distribution and, secondly, the univariate marginals have similar

heavy-tailed normal scale mixture distributions with (possibly) different tailedness parameters. As a consequence of practical interest, the DSNM allows for a different excess kurtosis on each dimension. We examine a number of properties of DSNMs such as the joint density function, hierarchical and stochastic representations, relations with other families of symmetric distributions, type of symmetry, univariate marginal distributions, no correlation implying independence, and moments of practical interest. For illustrative purposes, we describe two members of the DSNM family obtained in the case of components of the mixing random vector being either uniform or shifted exponential. These are examples of mixing distributions that guarantee a closed-form expression for the joint density of the DSNM. For the two DSNMs analyzed in detail, we introduce parsimony by allowing the d tailedness parameters to be tied across dimensions, and describe algorithms, based on the expectation-maximization (EM) principle, to estimate the parameters by maximum likelihood. We use real data from the financial and biometrical fields to appreciate the advantages of our DSNMs over other symmetric heavy-tailed distributions available in the literature.

(4) Agustin Mayo-Iscar, Universidad De Valladolid
Title: *Robust Proposals for Clustering Based on Trimming and Constraints*
Abstract: In the last twenty five years trimming and constraints based robust proposals for maximum likelihood estimation in model based clustering have been successfully developed. Proofs of consistency and evidences of a non negligible breakdown point are available for these procedures. There remains an open issue when applying estimators based on the joint application of trimming and constraints related to choosing the number of clusters, the level of trimming and the strength of the constraints. Exploratory tools for assisting the users in selecting them are available. A novel parametric boostrap based likelihood ratio test procedure has been developed for identifying input-parameter combinations linked to interesting clustering solutions. Statistical properties of this proposal and empirical evidences about its performance when applied to artificial and real data, containing contaminating observations, have been obtained.

**Session** 25: ***Topics in Design of Clinical Trials***
Organizer and Chair: Xuekui Zhang, University of Victoria
**Room: MB251, Time: 1:20 PM – 3:00 PM**

(1) Xuekui Zhang, University of Victoria
Title: *Pplication of Group Sequential Methods to the 2-in-1 Design and Its Extensions for Interim Monitoring*
Abstract: The 2-in-1 adaptive design allows seamless expansion of an ongoing Phase 2 trial into a Phase 3 trial to expedite a drug development program. Under a mild assumption that is expected to generally hold in practice, as guaranteed by Slepian's lemma, the trial can be tested at the full alpha level with or without expansion, scarifying no statistical power. The endpoint used for expansion decisions can be the same as or different from the primary endpoints, and there is no restriction on the expansion bar. Due to its flexibility and robustness, it has

drawn immediate attention to academic researchers and industry practitioners. The design has been substantially extended in the literature and successfully implemented in multiple trials.

Group sequential methods are a cornerstone in trial monitoring. A preliminary investigation suggests that it can be applied to the 2-in-1 design without no formal mathematical proof. In this paper, we provide formal proof of its sufficient condition to fill the gap, which unlocks the full potential of the 2-in-1 design and paves the way for its wider applications. In practice, we can verify the condition using simulation studies as suggested by FDA guideline document. We also discussed a special case and when it is guaranteed without simulation checking.

(2) Leilei Zeng, University of Waterloo
Title: *Design of Longitudinal Cluster Randomized Trials*
Abstract: Longitudinal cluster randomized trials are experiments in which intact social units (e.g. schools/clinics/families), rather than independent individuals, are randomized to intervention groups. When discrete time to event data is collected from such trials, the marginal discrete hazard can be expressed in terms of transition probability and the within cluster dependence structure is characterized via pairwise odds ratios. Transition models can be used for estimation and inference. We consider the design of longitudinal cluster randomized trials directed at evaluating the effectiveness of intervention on the event occurrence for a desired power. In particular, formula for number of required clusters is derived based on the robust variance estimators from the transition models that account for the intra-cluster associations. Simulations and an application on the data from Waterloo Smoking Prevention Project are used for illustration purposes.

(3) Xikui Wang, University of Manitoba
Title: *A New Model of the Continual Reassessment Method for Phase I Clinical Trials*
Abstract: Phase I dose-finding clinical trials of new drugs are fundamentally important for the overall success of drug development. One key assumption for such dose-finding trials is that increasing the drug dosage increases its toxicity probability. The goal is to estimate the maximum tolerated dose (MTD), which is the highest dose with an acceptable level of toxicity. This objective is especially important for cytotoxic drugs against cancer, which potentially have strong or even lethal toxicity. The continual reassessment method (CRM) is a model-based design to estimate MTD and assumes a parametric probability function to depict the toxicity probability at each dosage. There are three classic functions used in the literature: power, logistic and hyperbolic tangent. We introduce a new function and compare its performance against the classic functions. This is a joint work with W. Zhang, P. Yang and S. Muthukumarana.

(4) Yanqing Yi, Memorial University of Newfoundland
Title: *A Markov Decision Process for Response Adaptive Designs*
Abstract: The randomized treatment allocation process in a response adaptive clinical trial is formulated as a stochastic sequential decision problem and an algorithm is proposed to approximate the optimal value under the average reward

criterion. When the information of previous treatment allocations and associated responses are summarized with sufficient statistics for unknown parameters, the decision process becomes a Markov process, on which a span-contractor operator is defined. It is proven that the average reward under the policy identified from the span-contractor operator converges almost surely to the optimal value. Numerical results wil be presented.

**Session** 26: ***Statistical Methods for Biomedical Data Science***
Organizer: Dehan Kong, University of Toronto
Chair: Yi Liu, University of Alberta
**Room: Elder Tom Crane Bear, Time: 1:20 PM – 3:00 PM**

(1) Gen Li, University of Michigan
Title: *Scalar-on-Tensor Regression with Incomplete Observations*
Abstract: Multivariate longitudinal data measured on a regular grid can be concisely represented as a three-way tensor array (i.e., sample-by-feature-by-time). Missing observations are commonly encountered in such data since not all samples are measured at every time point. The missing data impose significant challenges for statistical analysis. This work develops a novel scalar-on-tensor regression framework, called TRIO, which effectively leverages all available observations in a design tensor for accurate parameter estimation and prediction. We propose a parsimonious model for the design tensor and regression coefficient matrix and devise a computationally efficient algorithm to estimate model parameters with flexible regularization. Numerical studies demonstrate the superior performance of the proposed method over competitors. The application to real data examples further provides novel insights.

(2) Yuying Xie, Michigan State University
Title: *Supervised Capacity Preserving Mapping: A Clustering Guided Visualization Method for scRNAseq Data*
Abstract: The rapid development of scRNA-seq technologies enables us to explore the transcriptome at the cell level in a large scale. Recently, various computational methods have been developed to analyze the scR-NAseq data such as clustering and visualization. However, current visualization methods including t-SNE and UMAP are challenged by the limited accuracy of rendering the geometic relationship of populations with distinct functional states. Most visualization methods are unsupervised, leaving out information from the clustering results or given labels. This leads to the inaccurate depiction of the distances between the bona fide functional states and the variance of clusters. We present supCPM, a robust supervised visualization method, which separates different clusters, preserves global structure, and tracks the cluster variance. Compared with six visualization methods using synthetic and real data sets, supCPM shows improved performance than other methods in preserving the global geometric structure and data variance. Overall, supCPM provides an enhanced visualization pipeline to assist the interpretation of functional transition and accurately depict population segregation

(3) Todd Ogden, Columbia University
Title: *Nonparametric Functional Data Modeling of Pharmacokinetic Processes with Applications in Dynamic PET Imaging*

Abstract: Modeling a pharmacokinetic process typically involves solving a system of linear differential equations and estimating the parameters upon which the functions depend. In order for this approach to be valid, it is necessary that a number of fairly strong assumptions hold, assumptions involving various aspects of the kinetic behavior of the substance being studied. In many situations, such models are understood to be simplifications of the "true" kinetic process. While in some circumstances such a simplified model may be a useful (and close) approximation to the truth, in some cases, important aspects of the kinetic behavior cannot be represented. We present a nonparametric approach, based on principles of functional data analysis, to modeling of pharmacokinetic data. We illustrate its use through application to data from a dynamic PET imaging study of the human brain.

(4) Lily Wang, George Mason University
Title: *An Efficient Spline Smoothing for 3D Point Cloud Learning*

Abstract: Over the past two decades, we have seen an exponentially increased amount of point clouds of irregular shapes collected in various areas. Motivated by the importance of solid modeling for point clouds, we develop a novel and efficient smoothing tool based on multivariate splines over the tetrahedral partitions to extract the underlying signal and build up a 3D solid model from the point cloud. The proposed method can be used to denoise or deblur the point cloud effectively and provide a multi-resolution reconstruction of the actual signal. In addition, it can handle sparse and irregularly distributed point clouds and recover the underlying trajectory from globally and locally missing data. Furthermore, we establish the theoretical guarantees of the proposed method. Specifically, we derive the convergence rate and asymptotic normality of the proposed estimator and illustrate that the convergence rate achieves the optimal nonparametric convergence rate, and the asymptotic normality holds uniformly. We demonstrate the efficacy of the proposed method over traditional smoothing methods through extensive simulation examples.

## Coffee Break, MB Central Foyer

# Parallel Sessions H
# 3:20pm - 5:00pm, Saturday, July 9th

**Session** 27: ***Challenges and Developments in New Data Era***
Organizer: Linglong Kong, University of Alberta
Chair: Wei Tu, Queen's University
**Room: MB Auditorium, Time: 3:20 PM – 5:00 PM**

(1) Jinhan Xie, University of Alberta
Title: *Statistical Inference for Smoothed Quantile Regression with Streaming*

*Data*

Abstract: In this paper, we address the problem of how to conduct the valid statistical inference for quantile regression with streaming data. The main difficulties are that the quantile regression loss function is non-smooth and it is often infeasible to store the entire dataset in memory, which invalidates the use of the existing methodology. To overcome these issues, we propose a fully online updating method for statistical inference in smoothed quantile regression with streaming data. Our main contributions are two-fold. First, in the low-dimensional regime, we present an incremental updating algorithm to obtain the smoothed quantile regression estimator with streaming data set. The proposed estimator allows us to construct asymptotically exact statistical inference procedures. Second, in the high-dimension regime, we develop an online debiased lasso procedure to accommodate the special sparse structure of streaming data. The proposed online debiased procedure is updated with only the current data and summary statistics of historical data and corrects an approximation error term from online updating with streaming data. Moreover, theoretical results such as estimation consistency and asymptotic normality are established to justify its validity in both settings. Simulation studies with supportive evidence are presented. Applications are illustrated with the Seoul bike sharing demand data and the appliances energy data.

(2) Wendy Lou, University of Toronto
Title: *An Integrated Approach with Multiple Longitudinal Markers for Disease Phenotyping*

Abstract: Motivated by two cohort studies with heterogeneous participant characteristics, we will present a joint modeling approach incorporating multiple co-existing longitudinal patterns to identify underlying phenotypes of the study population. The advantages of the proposed approach and the remaining challenges will be discussed via real examples of longitudinal markers with distinct and complex structures. The performance of the proposed method in comparison to existing models will also be evaluated via simulation studies under several scenarios. (This is a joint work with Zihang Lu at Queen's University).

(3) Faming Liang, Purdue University
Title: *Statistical Inference with Sparse Deep Learning*

Abstract: Deep learning has powered recent successes of artificial intelligence (AI). However, how to perform statistical inference with deep neural networks remains still an unresolved issue. We address this issue via spars deep learning. In particular, we lay down a theoretical foundation for sparse deep learning and propose some efficient algorithms for learning sparse neural networks. The former has successfully tamed the sparse deep neural network into the framework of statistical modeling, enabling relevant variables consistently identified and prediction uncertainty correctly quantified. The latter can be asymptotically guaranteed to converge to the global optimum, enabling the validity of the downstream statistical inference. Numerical result indicates validity of the proposed methods. The presentation is based on joint works with Yan Sun and Qifan Song

at Purdue University.

(4) Jian Kang, University of Michigan
Title: *Bayesian Spatially Varying Weight Neural Networks with the Soft-Thresholded Gaussian Process Prior*

Abstract: Deep neural networks (DNN) have been adopted in the scalar-on-image regression which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve a good prediction accuracy and the model fitting results can be difficult to interpret. In this work, we construct a novel single-layer Bayesian neural network (BNN) with spatially varying weights for the scalar-on-image regression. Our goal is to select interpretable image regions and to achieve high prediction accuracy with limited training samples. We assign the soft-thresholded Gaussian process (STGP) prior to the spatially varying weights and develop an efficient posterior computation algorithm based on stochastic gradient Langevin dynamics (SGLD). The BNN-STGP provides large prior support for sparse, piecewise-smooth, and continuous spatially varying weight functions, enabling efficient posterior inference on image region selection and automatically determining the network structures. We establish the posterior consistency of model parameters and selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. We compared our methods with state-of-the-art deep learning methods via analyses of multiple real data sets including the task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study.

**Session** 28: ***Non-Parametric Methods for Biomedical Data***
Organizer and Chair: Zhengwu Zhang, University of North Carolina at Chapel Hill
**Room: Elder Tom Crane Bear, Time: 3:20 PM – 5:00 PM**

(1) Benjamin Risk, Emory University
Title: *Correcting Sampling Bias in Neuroimaging Studies using Doubly Robust Nonparametric Inference*

Abstract: Neuroimaging studies remove participants that fail motion quality control criteria. Motion is particularly problematic in studies on children and neurodevelopmental disorders, including autism spectrum disorder (ASD). Popular motion quality control criteria result in the removal of participants with more severe ASD. To address the sampling bias, we define a target parameter for the difference in brain connectivity between ASD and typically developing children. We estimate the target parameter using doubly robust targeted minimum loss-based estimation (DRTMLE) with an ensemble of machine learning methods for the propensity and outcome models. In a study of four hundred children, we find more extensive differences than the naive estimator. Our approach can be used to reveal the pathophysiology of neurological disorders in populations with sampling bias.

(2) Yi Zhao, Indiana University
Title: *Hierarchical Tree Data in Regularized Regression: A Path Analysis Perspective*

Abstract: Brain segmentation at different levels is generally represented as hierarchical trees. Brain regional atrophy at specific levels was found to be marginally associated with Alzheimer's disease outcomes. In this study, we propose an l1-type regularization for predictors that follow a hierarchical tree structure. Considering a tree as a directed acyclic graph, we interpret the model parameters from a path analysis perspective. Under this concept, the proposed penalty regulates the total effect of each predictor on the outcome. With regularity conditions, it is shown that under the proposed regularization, the estimator of the model coefficient is consistent in l2-norm and the model selection is also consistent. When applied to a brain structural magnetic resonance imaging dataset acquired from the Alzheimer's Disease Neuroimaging Initiative, the proposed approach identifies brain regions where atrophy in these regions demonstrates the declination in memory. With regularization on the total effects, the findings suggest that the impact of atrophy on memory deficits is localized from small brain regions, but at various levels of brain segmentation.

(3) Fei Gao, Fred Hutchinson Cancer Center
Title: *Noniterative Adjustment to Regression Estimators with Population-Based Auxiliary Information for Semiparametric Models*

Abstract: Disease registries, surveillance data, and other datasets with extremely large sample sizes become increasingly available in providing population-based information on disease incidence, survival probability, or other important public health characteristics. Such information can be leveraged in studies that collect detailed measurements but with smaller sample sizes. In contrast to recent proposals that formulate additional information as constraints in optimization problems, we develop a general framework to construct simple estimators that update the usual regression estimators with some functionals of data that incorporate the additional information. We consider general settings that incorporate nuisance parameters in the auxiliary information, non-i.i.d. data such as those from case-control studies, and semiparametric models with infinite-dimensional parameters common in survival analysis. Details of several important data and sampling settings are provided with numerical examples.

(4) Zhengwu Zhang, University of North Carolina Chapel Hill
Title: *Analyzing Brain Structural Connectivity as Continuous Functions*

Abstract: Brain structural networks are often represented as discrete adjacency matrices, where each element in the matrix provides a summary of the connectivity between pairs of regions of interest (ROIs). These ROIs are typically determined a-priori using a brain atlas; a parcellation of the cortical surface constructed from anatomical considerations. Unfortunately, the choice of atlas is often arbitrary and can lead to a loss of important connectivity information at the sub-ROI level. This work introduces an atlas-independent framework that overcomes these issues by modeling brain connectivity using smooth functions. In particular, our framework assumes that the pattern of observed white matter fiber tract endpoints is driven by a latent random function defined over a product manifold domain, referred to as the continuous connectivity. As a result, our

framework is inherently both atlas and resolution-independent, and so prevents information loss caused by large ROIs. Under this continuous connectivity representation, we develop connectome alignment algorithms and statistical analysis frameworks to analyze brain network.

**Session** 29: ***Epidemic Modelling and Surveillance***
Organizer and Chair: Rob Deardon, Univeristy of Calgary
**Room: MB253, Time: 3:20 PM – 5:00 PM**

(1) Laura Cowen, University of Victoria
Title: *Estimating the Scope of the COVID-19 Pandemic in Canada.*
Abstract: In Canada, we have largely understood the COVID-19 pandemic through daily case counts. However, the true scope of the pandemic is unknown due to hidden or undetected cases. We estimate the pandemic scope through a new multi-site model using publicly available count data such as observed cases, recoveries among observed cases, and deaths. The model estimates the total number of active COVID-19 cases per region for each reporting interval (such as each week). We applied this multi-site model to Canada as a whole, with each province and territory acting as an individual site. The Canada model estimates the total COVID-19 burden for 90 weeks from 2020-04-02 to 2022-02-10. We also apply the multi-site model to the five Health Authority regions of British Columbia, Canada. We obtain simultaneous estimates for all five regions to produce an account of the pandemic over 31 weeks, starting 2020-04-02 and ending 2020-10-30.

(2) MD Mahsin, University of Calgary
Title: *Spatial Modeling of Infectious Disease Transmission using Continuous-Time Geographically-Dependent Individual-Level Mode*
Abstract: Modeling infectious diseases has been increasingly used to evaluate the potential impact of different control measures and guide public health policy decisions. There has been rapid progress in developing spatio-temporal modeling of infectious diseases in recent years. An example of recent developments is the discrete-time geographically-dependent individual-level models (GD-ILMs). A key feature of the GD-ILMs is that they allow for evaluating spatially varying risk factors, environmental factors, and unobserved spatial structure to account for geographical location upon infectious disease transmission. A conditional autoregressive model captures the effects of spatially structured latent covariates or measurement error. However, these models have been set in discrete-time and assumed known times of infection and removal and a constant infectious period. A more realistic approach is to build a model that considers infections and recoveries on a continuous-time scale and allows for censoring of the event times and heterogeneity of the infectious periods between individuals. Here, we propose a novel continuous-time GD-ILMs, allowing infection times and infectious periods as latent variables that are estimated using the data-augmented Markov Chain Monte Carlo techniques within a Bayesian framework. This approach results in a flexible infectious disease modeling framework for formulating etiological

hypotheses and identifying unusually high-risk geographical regions to develop preventive action. We evaluate the performance of these proposed models for simulated epidemic data and seasonal influenza data in Calgary in 2009.

(3) Madeline Ward, University of Calgary
Title: *Incorporating Behavioural Change into Spatial Individual-Level Models for Infectious Disease Transmission*

Abstract: Individual-level models can flexibly incorporate information on individual risk factors, including spatial location. This can account for the high degree of heterogeneity that is characteristic of population mixing, and, thus, infection transmission. However, these models have typically assumed stable population behaviour over time. As we have observed throughout the COVID-19 pandemic, behaviour often changes based on the current perceived risk of contracting the disease. In turn, this behaviour change can have a large impact on the transmission dynamics of the disease. We will present a new class of behavioural-change individual-level models where various functions of infection prevalence affect susceptibility and/or population mixing and illustrate their use through simulated and real data on the 2001 foot and mouth disease epidemic amongst livestock.

**Session** 30: ***Variable Selection Methods for Correlated Data in High-Dimensions***
Organizer: Sahir Bhatnagar, Mcgill University
Chair: Sahir Bhatnagar , Mcgill University
**Room: MB252, Time: 3:20 PM – 5:00 PM**

(1) Julien St-Pierre, McGill University
Title: *Efficient Penalized Generalized Linear Mixed Models for Variable Selection and Genetic Risk Prediction in High-Dimensional Data*

Abstract: In genome-wide association studies (GWAS), generalized linear mixed models (GLMMs) are now widely used to model population structure and/or cryptic relatedness by including a polygenic random effect with variance-covariance structure proportional to the genetic similarity matrix used to infer the Principal Components (PCs). Following a GWAS, a polygenic risk score (PRS) can be constructed by summing the risk alleles in an individual to obtain a single overall genetic risk. We propose to use a penalized quasi-likelihood loss function with a LASSO regularization to select important genetic predictors and estimate their effects, while controlling for population structure and/or cryptic relatedness by including a polygenic random effect, to derive a multivariate PRS for binary traits. We perform simulation studies to evaluate the performance of our proposed method in a variety of scenarios.

(2) Kevin McGregor, York University
Title: *Microbial Diversity Estimation and Hill Number Calculation using the Hierarchical Pitman-Yor Process*

Abstract: The human microbiome comprises the microorganisms that inhabit the human body. The composition of a microbial population is often quantified

through measures of species diversity, which summarize the number of species along with their relative abundances into a single value. In a microbiome sample there will certainly be species missing from the target population which will affect the diversity estimates. We employ a model based on the hierarchical Pitman-Yor (HPY) process to model the species abundance distributions over multiple populations. The model parameters are estimated using a Gibbs sampler. We also derive estimates of species diversity, conditional and unconditional on the observed data, as a function of the HPY parameters. Finally, we derive a general formula for the Hill numbers in the HPY context. We show that the Gibbs sampler for the HPY model performs well in simulations. We also show that the conditional estimates of diversity from the HPY model improve over naive estimates when species are missing. Similarly the conditional HPY estimates tend to perform better than the naive estimates especially when the number of individuals sampled from a population is small.

(3) Maxime Turgeon, University of Manitoba
Title: *Generalized Soft Impute for Matrix Completion*

Abstract: Missing data is a common challenge in data science. As the number of measurements increases, so does the likelihood that at least one of them is missing for a given observation, leading to inefficient complete-case analyses. Matrix completion algorithms have gained popularity recently for their simplicity and computational efficiency. In this talk, we present a matrix completion algorithm based on generalized Singular Value Decomposition (SVD), which unlike classical SVD imposes constraints on the rows and columns of the data matrix. This framework is particularly suitable for multivariate methods like Weighted Principal Component Analysis and Correspondence Analysis. We obtain good performance by regularizing the nuclear norm of the completed matrix, and we achieve computational efficiency by using proximal gradient descent. Finally, we discuss applications of our algorithm to the field of statistical genetics.

(4) Sahir Bhatnagar, McGill University
Title: *Variable Selection in Parametric Hazard Models*

Abstract: The semiparametric Cox model has become the default approach to survival analysis, even though Cox himself later suggested he would prefer to model the hazard function directly to do things like predict the outcome for a particular patient. Methods relying on time matching or risk-set sampling require a separate estimation of the baseline hazard for survival or cumulative incidence curves. Extending these methods to more complex settings, such as penalized regression, require specialized implementations. In this talk, we first introduce case-base sampling; a parametric approach where hazard functions can be estimated in continuous-time using logistic regression. This approach naturally leads to estimates of the survival or risk functions that are smooth-in-time. We then show how case-base sampling can be used for variable selection through regularized estimation of the hazard function. We contrast our approach with Coxnet, which regularizes the Cox partial likelihood.

**Session** 31: ***Statistical Learning in Large-Scale Medical Imaging Studies***
Organizer: Chao Huang, Florida State University
Chair: Junhao Zhu, University of Toronto
**Room: MB252, Time: 3:20 PM – 5:00 PM**

(1) Yafei Wang, University of Alberta
Title: *Bayesian Distributionally Robust Optimization with Discrete Finite Support*
Abstract: Distributionally Robust Optimization (DRO) has been recently considered as a principled approach to data-driven decision making problems owning to its state-of-art performance and robustness against the distribution drift. We propose a Bayesian framework for assessing the relative strengths of distributional robustness for the decision making problem under the scenario that the underlying distribution is defined by a finite-dimensional parameter. In general, the true underlying population distribution is rarely known but can be observed through finite training samples. The frequentist DRO considers using empirical distribution as the approximation of the true distribution, which ignores the prior information hidden in the data. By contrast, the key idea we propose in this paper is to approximate the true underlying distribution by predictive posterior. Moreover, we construct new ambiguity sets for constraints that are tractable and enjoy the Bayesian robustness guarantee. We establish that asymptotically, solutions to DRO models with our Bayesian framework enjoy strong robustness properties and performance guarantees, such as asymptotic consistency and finite sample guarantee. While much research effort has been devoted to tractable reformulations for DRO problems, due to the number of variables involved, few efficient numerical algorithms are developed, and most of them can neither handle the large-scale scenarios nor non-smooth loss function effectively. We fill the gap by reformulating the DRO as a sequence block dual problem and using sGS-ADMM that can achieve O(1/sqrt(N)) iteration complexity where the major computations involved in each iteration can be conducted in parallel. Finally, we also provide guidelines to select among competing ambiguity set in DRO for practitioner including moment-based, Wasserstein-based based, and f-divergence-based ambiguity set. The proposed method is evaluated through simulated data and portfolio allocation data.

(2) Hai Shu, New York University
Title: *Orthogonal Common-Source and Distinctive-Source Decomposition Between High-Dimensional Data Views*
Abstract: Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of two high-dimensional data views/sets is to decompose each data matrix into three parts: a low-rank common-source matrix that captures the shared information across data views, a low-rank distinctive-source matrix that characterizes the individual information within each single data view, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common-source and distinctive-source matrices, but

inadequately consider the more necessary orthogonal relationship between the two distinctive-source matrices. The latter guarantees that no more shared information is extractable from the distinctive-source matrices. We propose a novel decomposition method that defines the common-source and distinctive-source matrices from the L2 space of random variables rather than the conventionally used Euclidean space, with a careful construction of the orthogonal relationship between distinctive-source matrices. The proposed estimators of common-source and distinctive-source matrices are shown to be asymptotically consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and the real data analysis.

(3) Xiaoxiao Li, The University of British Columbia
Title: *Learning Neuroimaging Data with Deep Graph Neural Network*

Abstract: The brain is an exceptionally complex system. Using noninvasive medical neuroimaging to understand its functional organization is the goal of modern neuroscience. Numerous significant techniques for mapping the structural and functional connectivity of the brain have been developed in order to create a comprehensive road map of neuronal activity in the human brain. Large strides in understanding the human brain have been made by modeling individual brain or populational neuroimaging as a graph — a mathematical construct describing the connections or interactions (i.e. edges) between different discrete objects (i.e. nodes). Graph Convolutional Neural Networks (GCNs) are widely used for graph analysis. However, like the other deep learning models, GCNs lack explainability and rely on a vast amount of data for training. To address these limitations, we have designed a novel interpretable GCN model, BrainGNN, for brain network analysis and developed a novel federated graph learning scheme, FedNI, and a provable self-supervised learning method, GATE, to improve GCNs' performance on the limited neuroimaging data collected in a single center. We conduct extensive experiments on datasets across cohorts and modalities. We demonstrate the power of GCN and our novel designs to advance the analysis of neuroimaging data.

(4) Dengdeng Yu, UNIVERSITY of TEXAS at ARLINGTON
Title: *Mapping the Genetic-Imaging-Clinical Pathway with Applications to Alzheimer's Disease*

Abstract: Alzheimer's disease is a progressive form of dementia that results in problems with memory, thinking, and behavior. It often starts with abnormal aggregation and deposition of beta-amyloid and tau, followed by neuronal damage such as atrophy of the hippocampi, leading to Alzheimer's Disease (AD). The aim of this paper is to map the genetic-imaging-clinical pathway for AD in order to delineate the genetically regulated brain changes that drive disease progression based on the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset. We develop a novel two-step approach to delineate the association between high-dimensional 2D hippocampal surface exposures and the Alzheimers Disease Assessment Scale (ADAS) cognitive score, while taking into account the ultra-high dimensional clinical and genetic covariates at baseline. Analysis

results suggest that the radial distance of each pixel of both hippocampi is negatively associated with the severity of behavioral deficits conditional on observed clinical and genetic covariates. These associations are stronger in Cornu Ammonis region 1 (CA1) and subiculum subregions compared to Cornu Ammonis region 2 (CA2) and Cornu Ammonis region 3 (CA3) subregions.

## 5:15-6:45 PM ICSA-Canada Chapter Annual General Meeting (AGM)
## Room: Elder Tom Crane Bear.

## 7:00-10:00 PM Banquet
## Room: Kinnear Centre 103-105

# Plenary Talk III
# 8:30am - 9:30am, Sunday, July 10th

**Session** 32: ***Keynote Speech 3***
Organizer: Yingwei Peng, Queen's University
Chair: Dehan Kong, University of Toronto
**Room: MB Auditorium, Time: 8:30 AM – 9:30 AM**

(1) Heping Zhang, Yale University
Title: *Tensor Quantile Regression for Neuroimage Study of Human Intelligence*
Abstract: Human intelligence is usually measured by well-established psychometric tests through a series of problem solving. The recorded cognitive scores are continuous but usually heavy-tailed with potential outliers and violating the normality assumption. Meanwhile, magnetic resonance imaging provides an unparalleled opportunity to study brain structures and cognitive ability. Motivated by association studies between MRI images and human intelligence, we propose a tensor quantile regression model, which is a general and robust alternative to the commonly used scalar-on-image linear regression. Moreover, we take into account rich spatial information of brain structures, incorporating low-rankness and piece-wise smoothness of imaging coefficients into a regularized regression framework. We formulate the optimization problem as a sequence of penalized quantile regressions with a generalized Lasso penalty based on tensor decomposition, and develop a computationally efficient alternating direction method of multipliers algorithm estimate the model components. Extensive numerical studies are conducted to examine the empirical performance of the proposed method and its competitors. Finally, we apply the proposed method to a large-scale important dataset: The Human Connectome Project. We find that the tensor quantile regression can serve as a prognostic tool to assess future risk of cognitive impairment progression. More importantly, with the proposed method, we are able to identify the most activated brain subregions associated with quantiles of human intelligence. The prefrontal and anterior cingulate cortex are found to be mostly associated with lower and upper quantile of fluid intelligence. The insular cortex associated with median of fluid intelligence is a rarely reported region. This is a joint work with Cai Li, Assistant Professor of Biostatistics, St. Jude Children's Hospital.

## Coffee Break, MB Central Foyer

# Parallel Sessions J
# 9:50am - 11:30pm, Sunday, July 10th

**Session** 33: ***Recent Advances in Functional Data Analysis***
Organizer and Chair: Peijun Sang, University of Waterloo
**Room: MB Auditorium, Time: 9:50 AM – 11:30 AM**

(1) Bing Li, Pennsylvania State University
Title: *Functional Directed Acyclic Graphs*

Abstract: We introduce a new method to estimate directed acyclic graphs from multivariate functional data, based on the notion of faithfulness that relates a directed acyclic graph with a set of conditional independence relations among the random functions. To characterize and evaluate these relations, we propose two linear operators, the conditional covariance operator and the partial correlation operator. Based on these operators, we adapt and extend the PC-algorithm to estimate the functional directed graph, so that the computation time depends on the sparsity rather than the full size of the graph. We study the asymptotic properties of the two operators, derive their uniform convergence rates, and establish the uniform consistency of the estimated graph, all of which are obtained while allowing the graph size to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and an application to a time-course proteomic dataset.

(2) Fang Yao, Peking University
Title: *Online Estimation for Functional Data*

Abstract: Functional data analysis has attracted considerable interest, and is facing new challenges of the increasingly available data in streaming manner. In this work, we propose a new online method to dynamically update the local linear estimates of mean and covariance functions of functional data, which is the foundation of subsequent analysis. The kernel-type estimates can be decomposed into two sufficient statistics depending on the data-driven bandwidths. We propose to approximate the future optimal bandwidths by a dynamic sequence of candidates and combine the corresponding statistics across blocks to make an updated estimation. The proposed online method is easy to compute based on the stored sufficient statistics and current data block. Based on the asymptotic normality of the online mean and covariance function estimates, the relative efficiency in terms of integrated mean squared error is studied and a theoretical lower bound is obtained. This bound provides insight into the relationship between estimation accuracy and computational cost driven by the length of candidate bandwidth sequence that is pivotal in the online algorithm. Simulations and real data applications are provided to support such findings and show the advantages of the proposed method.

(3) Yehua Li, University of California, Riverside
Title: *Semiparametric Functional Regression Models with Multivariate Functional Predictors*

Abstract: Motivated by an application on predicting crop yield using temperature trajectories and other scalar predictors, we consider two classes of semiparametric functional regression models, both of which are extensions of the classic functional linear models. We jointly model multiple functional predictors that are cross-correlated using multivariate functional principal component analysis (mFPCA), and use the mFPCA score as extracted features in a second stage semiparametric regression. In the proposed partially linear functional additive

models (PLFAM), we predict the scalar response by both the parametric effects of the multivariate predictor and additive nonparametric effects of the mFPCA scores, and adopt the component selection and smoothing operator (COSSO) penalty to select relevant components and regularize the fitting. In the second class of semiparametric functional regression models, we also consider the interactions between the functional and multivariate predictors, where we assume the interaction depends on a nonparametric, single-index structure of the multivariate predictor to avoid the curse of dimensionality. We establish theoretic properties for both models, where we let the number of principal components diverge to infinity with the sample size. A fundamental difference between our framework and the existing high-dimensional semiparametric regression models is that the principal component scores are estimated with errors, the magnitudes of which increase with the order of FPC. The practical performances of the proposed methods are illustrated through analysis of the motivating crop yield data.

(4) Zuofeng Shang, New Jersey Institute of Technology
Title: *Deep Neural Network Classifier for Multi-Dimensional Functional Data*
Abstract: We propose a new approach, called as functional deep neural network (FDNN), for classifying multi-dimensional functional data. Specifically, a deep neural network is trained based on the principle components of the training data which shall be used to predict the class label of a future data function. Unlike the popular functional discriminant analysis approaches which rely on Gaussian assumption, the proposed FDNN approach applies to general non-Gaussian multi-dimensional functional data. Moreover, when the log density ratio possesses a locally connected functional modular structure, we show that FDNN achieves minimax optimality. The superiority of our approach is demonstrated through both simulated and real-world datasets.

**Session** 34: ***Recent Development in Causal Inference***
Organizer and Chair: Yeying Zhu, University of Waterloo
**Room: MB251, Time: 9:50 AM – 11:30 AM**

(1) Mireille Schnitzer, Université De Montréal
Title: *Estimands and Estimation of COVID-19 Vaccine Effectiveness under the Test-Negative Design: Connections to Causal Inference*
Abstract: The test-negative design (TND) is routinely used for the monitoring of seasonal flu vaccine effectiveness. More recently, it has become integral to the estimation of COVID-19 vaccine effectiveness, in particular for more severe disease outcomes. Distinct from the case-control study, the design typically involves recruitment of participants with a common symptom presentation who are being tested for the infectious disease in question. Participants who test positive for the target infection are the "cases" and those who test negative are the "controls". Logistic regression is the only statistical method that has been proposed to estimate vaccine effectiveness under the TND while adjusting for confounders. While under strong modeling assumptions it produces estimates of a causal risk

ratio, it may be biased in the presence of effect modification by a confounder. I will present and justify an inverse probability of treatment weighting (IPTW) estimator for the marginal risk ratio, which is valid under effect modification. I'll discuss connections between the estimands targeted by these two methods and causal parameters under different interference assumptions. I will then describe the results of a simulation study to illustrate and confirm the derivations and to evaluate the performance of the estimators.

(2) Karim Mohammad Ehsanul, The University of British Columbia
Title: *Finite Sample Properties of Inverse Probability of Adherence Weighted Estimator of the per-Protocol Effect for Sustained Treatment Strategies*

Abstract: We (Mosquera, Karim, Hossain) investigated the finite sample performances of Inverse Probability (of Adherence) Weighted per-protocol (IPW-PP) estimators to address medication non-adherence in the context of a pragmatic randomized controlled trial. We compared the performances of IPW-PP estimators with commonly used naive and baseline adjusted per-protocol estimators, under different data-generating mechanisms (DGMs) emulating pragmatic trials, comparing two sustained treatment strategies, possibly with a non-null effect. DGMs include (i) different roles of a baseline variable; whether future time-varying prognostic factors are impacted by past adherence; and whether the baseline variable is measured, (ii) whether adherence patterns observed in two arms are differential, and when we have access to measurements of adherence and confounders that are recorded infrequently (sparsely). When baseline confounders are adjusted, we generally obtain unbiased estimates, but if some necessary variables are not measured, the IPW-PP estimator may still be preferable. High non-adherence patterns might negatively impact IPW-PP effect estimators, particularly when DGMs include confounding that may be impacted by previous adherence history. Further bias may be introduced when post-baseline measurements of adherence are recorded infrequently. We used the above estimators to analyze a case study from the Lipid Research Clinics Coronary Primary Prevention Trial data in the presence of non-adherence.

(3) Zhaohan Sun, University of Waterloo
Title: *Estimation of Network Treatment Effects with Nonignorable Missing Confounders*

Abstract: In causal inference literature, interference happens when the intervention on one unit affects the outcome of other units. Most of the previous methods for estimating the network causal effects assume that the covariates information are complete, which may lead to biased estimates when the missingness exists. In this study, we consider the partial interference setting, that is, the whole population can be partitioned into clusters where the outcome of each unit depends on the intervention on other units within the same cluster, but not on the units in different clusters. We also assume that the confounders are subject to nonignorable missingness. We propose three unbiased estimators for the direct, indirect, total, and overall effect of the intervention on the outcome, and derive the asymptotic results accordingly. A comprehensive simulation study is carried

out as well to investigate the finite sample properties of the proposed estimators. We illustrate the proposed methods by analyzing the dataset collected from an Acid Rain Program, which was launched to reduce air pollution in the USA by encouraging the scrubber's installation on power plants, where the records of some operating characteristics of the power generating facilities are subject to missingness.

(4) Shujie Ma, University of California, Riverside
Title: *Causal Inference via Artificial Neural Networks: From Prediction to Causation*

Abstract: Recent technological advances have created numerous large-scale datasets in observational studies, which provide unprecedented opportunities for evaluating the effectiveness of various treatments. Meanwhile, the complex nature of large-scale observational data post great challenges to the existing conventional methods for causality analysis. In this talk, I will introduce a new unified approach that we have proposed for efficiently estimating and inferring causal effects using artificial neural networks. We develop a generalized optimization estimation through moment constraints with the nuisance functions approximated by artificial neural networks. This general optimization framework includes the average, quantile and asymmetric least squares treatment effects as special cases. The proposed methods take full advantage of the large sample size of large-scale data and provide effective protection against mis-specification bias while achieving dimensionality reduction. We also show that the resulting treatment effect estimators are supported by reliable statistical properties that are important for conducting causal inference.

**Session** 35: ***Optimal Sampling Designs for Analyses of Microbiome and Anthropometry Data***
Organizer and Chair: Ying Zhang, Acadia University
**Room: MB253, Time: 9:50 AM – 11:30 AM**

(1) Hong Gu, Dalhousie University
Title: *Optimal Sampling Schemes for Modelling Microbiome Temporal Dynamics using an OU Process*

Abstract: The temporal dynamics of the microbiome have recently become an area of great interest. Most research has observed the stability and mean reversion for some microbiomes. However, little has been done to study the mean reversion rates of these stable microbes and how sampling frequencies are related to such conclusions.

We focus on the temporal dynamics of individual genera, absorbing all interactions in a stochastic term. We use simple stochastic differential equation models to assess the following three questions. (1) Does the microbiome exhibit temporal continuity? (2) Does the microbiome have a stable state? (3) To better understand the temporal dynamics, how frequently should data be sampled in future studies?

We find that a simple Ornstein-Uhlenbeck (OU) model which incorporates both temporal continuity and reversion to a stable state fits the data for all genera better than a Brownian motion model that contains only temporal continuity. The OU model also fits the data better than modelling separate time points as independent. Under the OU model, we calculate the variance of the estimated mean reversion rate (the speed with which each genus returns to its stable state). Based on this calculation, we are able to determine the optimal sample schemes for studying temporal dynamics.

(2) Toby Kenney, Dalhousie University
Title: *Sampling Schemes for OU Models of Microbiome Data with Measurement Error.*

Abstract: The temporal dynamics of the microbiome have recently become an area of great interest. Evidence suggests that microbial systems have some temporal persistence and some mean reversion. However, the time scales of these processes are largely unknown. Preliminary analyses suggest that for the human microbiome, the time-scale may have a half-life in the order of 1–2 days. However, these analyses did not account for the measurement error in the microbiome, which could impact the estimates.

An important question is how often to sequence the data. In recent work (presented in another talk) we solved the problem for a pure OU process, based on the information matrix. However it is possible that the optimal sampling scheme will change when we include measurement error in the models.

In this talk, I will extend the work on modelling the microbiome, and on Fisher information for an OU process to include measurement error. The Fisher information is more complicated, so finding the optimal sampling scheme is more challenging. I will present our findings on the estimated mean reversion rates and optimal sampling schemes under these conditions.

(3) Wilson Lu, Acadia University
Title: *Multivariate Probability Proportional to Size Sampling Design on Anthropometric Data*

Abstract: The multivariate probability proportional to size (MPPS) sampling design is primarily used to tackle multiple objective sampling problems in agriculture survey. In this talk, we use a simulation study to demonstrate the use of MPPS as a potentially optimal sampling design to anthropometry. We also apply this MPPS technique with some modifications to a large anthropometric data set from China. Our results show that MPPS has promising potential in anthropometrics survey and other real world applications.

(4) Xiaojian Xu, Brock University
Title: *Optimal Designs for Generalized Linear Mixed Models*

Abstract: We develop the methods of constructing the optimal sequential designs for generalized linear mixed models (GLMMs). The possible impact on the inference precision made when approximation appears in an assumed GLMM is investigated. The designs constructed are robust with protection on different

types of model departures. Both I-optimality and D-optimality are employed. A simulation study assesses the resulting I- and D-optimal designs in terms of integrated mean squared errors of the estimators for the parameters involved in the fixed effects in the link predictor (possibly misspecified). We conclude that the I-optimal designs outperform D-optimal designs for most of the cases considered, and both I- and D-optimal designs proposed are more efficient than the conventionally used uniform designs and the classical D-optimal designs obtained when assuming the fitted GLMM is precise. Although the design problems in a general setting of GLMMs are addressed, the commonly used logistic mixed models and application for microbiome studies are demonstrated.

**Session** 36: ***Statistical Methods for Complex Data***
Organizer: Yinli Qin, University of Waterloo
Chair: Liqun Diao, University of Waterloo
**Room: MB252, Time: 9:50 AM – 11:30 AM**

(1) Xiaowu Dai, University of California, Berkeley
Title: *Kernel Ordinary Differential Equations*
Abstract: Ordinary differential equation (ODE) is widely used in modeling biological and physical processes in science. In this article, we propose a new reproducing kernel-based approach for estimation and inference of ODE given noisy observations. We do not assume the functional forms in ODE to be known, or restrict them to be linear or additive, and we allow pairwise interactions. We perform sparse estimation to select individual functionals, and construct confidence intervals for the estimated signal trajectories. We establish the estimation optimality and selection consistency of kernel ODE under both the low-dimensional and high-dimensional settings, where the number of unknown functionals can be smaller or larger than the sample size. Our proposal builds upon the smoothing spline analysis of variance (SS-ANOVA) framework, but tackles several important problems that are not yet fully addressed, and thus extends the scope of existing SS-ANOVA as well. We demonstrate the efficacy of our method through numerous ODE examples.

(2) Hao Chen, University of California, Davis
Title: *A Universal Nonparametric Event Detection Framework for Modern Data*
Abstract: After observing snapshots of a network, can we tell if there has been a change in dynamics? After collecting spiking activities of thousands of neurons in the brain, how shall we extract meaningful information from the recording? We introduce a change-point analysis framework utilizing graphs representing the similarity among observations. This approach is non-parametric and can be applied to data when an informative similarity measure can be defined. Analytic approximations to the significance of the test statistics are derived to make the method fast applicable to long sequences. The method is illustrated through the analysis of the Neuropixels data.

(3) Liqun Diao, University of Waterloo
Title: *Adaptive Response-Dependent Two-Phase Designs: Some Results on Ro-*

*bustness and Efficiency*

Abstract: Large cohort studies routinely create biobanks in which biospecimens are stored for use in future biomarker studies. In such settings, two-phase response-dependent sampling designs involve sub-sampling individuals in the cohort, assaying their biospecimen to measure an expensive biomarker, and using this data to estimate key parameters of interest under budgetary constraints. When analyses are based on inverse probability weighted estimating functions, recent work has described adaptive two-phase designs in which a preliminary phase of sub-sampling based on a standard design facilitates approximation of an optimal selection model for a second sub-sampling phase. In this paper, we refine the definition of an optimal sub-sampling scheme within the framework of adaptive two-phase designs, describe how adaptive two-phase designs can be used when analyses are based on the likelihood or conditional likelihood, and consider the setting of a continuous biomarker where the nuisance covariate distribution is estimated nonparametrically at the design stage and analysis stage as required; efficiency and robustness issues are investigated. We also explore these methods for the surrogate variable problem and describe a generalization to accommodate multiple stages of phase II sub-sampling. A study involving individuals with psoriatic arthritis is considered for illustration, where the aim is to assess the association between the biomarker MMP-3 and the development of joint damage.

(4) Mu Zhu, University of Waterloo
Title: *Some Statistical Applications of Generative Neural Networks*

Abstract: The "deep learning" movement started very much outside of mainstream statistics. How can we also benefit from this movement? I will share some of our rudimentary successes in using generative neural networks for quasi Monte Carlo and probability forecasts, based on joint work with Marius Hofert and Avinash Prasad.

**Session** 37: ***Statistical Methods for Extreme Value Analysis***
Organizer: Dehan Kong, University of Toronto
Chair: Dingke Tang, University of Toronto
**Room: Elder Tom Crane Bear, Time: 9:50 AM – 11:30 AM**

(1) Sebastian Engelke, University of Geneva
Title: *Extremal Graphical Models*

Abstract: Conditional independence, graphical models and sparsity are key notions for parsimonious models in high dimensions and for learning structural relationships in the data. The theory of multivariate and spatial extremes describes the risk of rare events through asymptotically justified limit models such as max-stable and multivariate Pareto distributions. Statistical modeling in this field has been limited to moderate dimensions so far, owing to complicated likelihoods and a lack of understanding of the underlying probabilistic structures.

In Engelke and Hitz (2020, JRSSB) we introduce a new notion of conditional independence for multivariate Pareto distributions that allows us to define ex-

tremal graphical models in a natural way. Statistical inference for such sparse models can be simplified to lower-dimensional margins. For a popular parametric class of multivariate Pareto distributions we show that, similarly to the Gaussian case, the sparsity pattern of a general graphical model can be easily read of from suitable inverse covariance matrices. We give an overview over existing results and present recent work on structural properties of these models. Moreover, we study extensions to more general graphical models on Poisson point processes, including connections to Lévy processes.

(2) Stanislav Volgushev, University of Toronto
Title: *Structure Learning for Extremes*
Abstract: Extremal graphical models are sparse statistical models for multivariate extreme events. The underlying graph encodes conditional independencies and enables a visual interpretation of the complex extremal dependence structure. For the important case of tree models, we provide a data-driven methodology for learning the graphical structure. We show that sample versions of the extremal correlation and a new summary statistic, which we call the extremal variogram, can be used as weights for a minimum spanning tree to consistently recover the true underlying tree. Remarkably, this implies that extremal tree models can be learned in a completely non-parametric fashion by using simple summary statistics and without the need to assume discrete distributions, existence of densities, or parametric models for marginal or bivariate distributions. Extensions to more general graphs are also discussed.

(3) Nicola Gnecco, University of Geneva
Title: *Extremal Random Forests*
Abstract: Classical methods for quantile regression fail in cases where the quantile of interest is extreme and only few or no training data points exceed it. Asymptotic results from extreme value theory can be used to extrapolate beyond the range of the data, and several approaches exist that use linear regression, kernel methods or generalized additive models. Most of these methods break down if the predictor space has more than a few dimensions or if the regression function of extreme quantiles is complex. We propose a method for extreme quantile regression that combines the flexibility of random forests with the theory of extrapolation. Our extremal random forest (ERF) estimates the parameters of a generalized Pareto distribution, conditional on the predictor vector, by maximizing a local likelihood with weights extracted from a quantile random forest. Under certain assumptions, we show consistency of the estimated parameters. Furthermore, we penalize the shape parameter in this likelihood to regularize its variability in the predictor space. Simulation studies show that our ERF outperforms both classical quantile regression methods and existing regression approaches from extreme value theory. We apply our methodology to extreme quantile prediction for U.S. wage data.

(4) Huixia Wang, George Washington University
Title: *Extreme Quantile Estimation Based on the Tail Single-Index Model*
Abstract: It is important to quantify and predict rare events that have significant

societal effects. Existing works on analyzing such events rely mainly on either inflexible parametric models or nonparametric models that are subject to the "curse of dimensionality." We propose a new semiparametric approach based on the tail single-index model to obtain a better balance between model flexibility and parsimony. The procedure involves three steps. First, we obtain a sqrt-n-estimator of the index parameter. Next, we apply the local polynomial regression to estimate the intermediate conditional quantiles. Lastly, these quantiles are extrapolated to the tails to estimate the extreme conditional quantiles. We establish the asymptotic properties of the proposed estimators. Furthermore, we demonstrate using a simulation and an analysis of Los Angeles mortality and air pollution data that the proposed method is easy to compute and leads to more stable and accurate estimations than those of alternative methods.

**Lunch, 11:30 AM - 1:20 PM, Vistas Dining Room**

# Parallel Sessions K
# 1:20pm - 3:00pm, Sunday, July 10th

**Session** 38: ***Approaches to Measurement Error and Misclassification***
Organizer and Chair: Hua Shen, University of Calgary
**Room: MB Auditorium, Time: 1:20 PM – 3:00 PM**

(1) Liqun Wang, University of Manitoba
Title: *Instrumental Variable Estimation in Measurement Error Models with Ordinal Responses*

Abstract: Researchers in the medical, health, and social sciences routinely encounter ordinal data such as self-reports of health or happiness. When modelling ordinal outcome variables, it is common to have covariates (e.g., attitudes, family income, retrospective variables) that either cannot be measured directly or are measured with substantial error. It is well known that ignoring such error in covariates can bias coefficients and hence undermine correct estimates of their effects.

We propose an instrumental variable approach to estimation of a probit model with ordinal response and mismeasured predictor variables. We derive likelihood based and method of moments estimators which are consistent and asymptotically normally distributed under general conditions. This is achieved through linear projection of the unobserved true predictors onto the instrumental variables to create an auxiliary probit model with observed predictors. Our simulation studies show that these methods are effective in correction of bias caused by the measurement error. Moreover, the proposed estimators perform well in finite samples and are robust against the departure of normality assumption for the measurement error. Some real data applications further demonstrate the usefulness of the proposed method.

This is a joint work with J. Guan, K. Bollen and R. Thomas.

(2) Lang Wu, University of British Columbia
Title: *A Nonlinear Measurement Error Model for Survival Analysis*
Abstract: We consider measurement errors in time-dependent covariates of a survival regression model. To address measurement errors in the time-dependent covariates, we model the covariate process based on a nonlinear mixed effects model. Inference is then based on the (joint) likelihood of the covariate model and the survival model. We also discuss a multiple imputation method to address the measurement errors. The methods are applied to an HIV/AIDS dataset.

(3) Zheng Yu, University of Calgary
Title: *Analysis of Probability Sample and Non-Probability Sample Subject to Misclassification*
Abstract: We consider integrating a non-probability sample with misclassified response and a probability sample with relevant auxiliary information only. We propose a two-stage process to estimate the population mean of the response variable in the absence of validation data in a latent-variable framework. We first use an expectation-maximization algorithm based on the surrogate variables to enable the estimation of the unknown response and parameter vectors in regression model and misclassification model in the non-probability sample. We then utilize the estimates obtained to estimate the population mean of the response variable using different methods while focusing on the doubly robust estimators. The performance of the proposed method and its advantages over the naive and ad hoc methods are demonstrated in simulation studies. We apply the proposed method to a study on community's environmental well-being and quality of life.

**Session** 39: ***Statistical Learning and Causal Inference***
Organizer and Chair: Mireille Schnitzer, Université De Montréal
**Room: Elder Tom Crane Bear, Time: 1:20 PM – 3:00 PM**

(1) Kuan Liu, University of Toronto
Title: *Bayesian Approaches to Causal Inference with Latent Confounders*
Abstract: The strongly ignorable treatment assignment assumption (also known as no unmeasured confounding) requires a sufficiently large set of covariates being measured to ensure that subjects are exchangeable across the observed exposure given measured covariates. Although administrative data are rich in information, key confounders might not be captured. Several Bayesian sensitivity analyses for unmeasured confounding have been developed that use bias parameters to capture the effect of latent confounders on the outcome and exposure. However, there is a lack of considerations to handle time-dependent latent confounders. In this talk, I will present two parametric Bayesian causal approaches to tackle causal estimation when the causal structure features time-dependent latent confounders. We will discuss these methods under different simulated scenarios with varying strength of relations between the latent confounders and the observed variables.

(2) Janie Coulombe, McGill University
Title: *Causal Inference with Data Subject to Covariate-Dependent Observation Times: An Application to a Cohort of New Users of Antidepressants*

Abstract: Marginal and conditional treatment effects are often estimated using data from observational studies like those from electronic health records (EHR). These data contain rich, longitudinal information on treatments, outcomes, and potential treatment effect modifiers. Under a set of causal assumptions, they can be used to estimate causal treatment effects. In most observational studies, however, observation (or visit) times are not common across patients, which can affect the inference. In this talk, we discuss the issue of covariate-dependent observation times in causal inference in general. Then, focusing on the estimation of conditional treatment effects, we propose a new consistent estimator for optimal individualized treatment rules that can be used to tailor treatment to patient characteristics all while accounting for visit irregularity across patients. The method is applied to data from the Clinical Practice Research Datalink in the United Kingdom to build an optimal treatment rule that chooses between two commonly prescribed antidepressants, citalopram and fluoxetine. The aim of that rule is to minimize a detrimental weight change in patients with depression.

(3) Linbo Wang, University of Toronto
Title: *Fighting Noise with Noise: Mendelian Randomization with Pseudo Variables*

Abstract: Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method that constructs pseudo variables to identify and remove irrelevant candidate instruments having spurious correlations with the exposure. Theoretical and synthetic data analyses show that the proposed method performs favorably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

(4) Yongjin Park, University of British Columbia
Title: *Resolving Causality in Single-Cell Biology*

Abstract: Single-cell genomics has created great interest in fundamental biology and translational research in medicine. A typical sequencing platform is now powerful enough to measure tens of thousands of genes across hundreds of thousands of cells as a routine experiment. If the past ten to twenty years of genomics were devoted to finding disease genes and their regulatory circuits at a tissue level resolution, now biomedical researchers seek to find causal mechanisms at a cell level resolution, an ultimate unit of life.

Existing single-cell analysis methods tend to emphasize exploratory data analysis, optimizing for better visualization to appeal to non-quantitative scientists. Not undermining the success of manifold learning in single-cell biology, specifically tailored to answer biological questions, our group focuses on developing a complementary tool that enables causal interpretations of classical biostatistics models and methods in single-cell genomics.

This talk will introduce our efforts to identify causally-differentially-expressed genes in observational single-cell RNA-seq data. Our goal is to make Neyman-Rubin's potential outcome framework accessible to the bioinformatics community. We developed a scalable machine learning approach to handle a large number of cells in a modest computing environment.

**Session** 40: ***Topics in High-Dimensional Statistics***
Organizer: Yi Lian, Mcgill University
Chair: Yi Lian , Mcgill University
**Room: MB251, Time: 1:20 PM – 3:00 PM**

(1) Xiaomeng Ju, University of British Columbia
Title: *Robust Gradient Boosting for Regression Problems*

Abstract: When it comes to making predictions in non-parametric settings, boosting is highly flexible and scalable to data with many explanatory variables. In spite of its popularity and practical success, it is well-known that boosting may provide poor estimates when data have outliers. We present a two-stage boosting algorithm similar to what is done for robust linear MM-regression, which first minimizes a robust residual scale estimator, and then improves it by optimizing a bounded loss function. Unlike previous robust boosting proposals this approach does not require computing an ad hoc residual scale estimator in each boosting iteration. We address the issue of the initialization of our boosting algorithm and provide a permutation-based procedure to robustly measure the importance of each variable. The effectiveness of our method is illustrated using simulated and benchmark data in regular and high dimensional settings and it compares favourably to existing methods: with clean data, our method works equally well as gradient boosting with the squared loss; with symmetric and asymmetrically contaminated data, our proposal outperforms other boosting methods (robust or otherwise) in terms of prediction error.

(2) Zhiyu Quan, University of Illinois at Urbana-Champaign
Title: *Improving Business Insurance Loss Models by Leveraging InsurTech Innovation*

Abstract: Recent transformative and disruptive developments in the insurance industry embrace various InsurTech innovations. Particularly with the rapid advances in data science and computational infrastructure, InsurTech is able to incorporate multiple emerging sources of data and reveal implications for value creation on business insurance by enhancing current insurance operations. In this paper, we unprecedently combine real-life proprietary insurance claims information and features, empowered by InsurTech, describing insured businesses

to create enhanced tree-based loss models. Empirical study shows that the supplemental data sources created by InsurTech innovation help significantly improve the underlying insurance company's in-house or internal pricing models. We further demonstrate how InsurTech proliferates firm-level value creation and affect insurance product development, pricing, underwriting, claim management and administration practice.

(3) Ziang Niu, University of Pennsylvania
Title: *Estimation and Inference for High-Dimensional Nonparametric Additive Instrumental-Variables Regression*

Abstract: The method of instrumental variables provides a fundamental and practical tool for causal inference in many empirical studies where unmeasured confounding between the treatments and the outcome is present. Modern data such as the genetical genomics data from these studies are often high-dimensional. The high-dimensional linear instrumental-variables regression has been considered in the literature due to its simplicity albeit a true nonlinear relationship may exist. We propose a more data-driven approach by considering the nonparametric additive models between the instruments and the treatments while keeping a linear model between the treatments and the outcome so that the coefficients therein can directly bear causal interpretation. We provide a two-stage framework for estimation and inference under this more general setup. The group lasso regularization is first employed to select optimal instruments from the high-dimensional additive models, and the outcome variable is then regressed on the fitted values from the additive models to identify and estimate important treatment effects. We provide non-asymptotic analysis of the estimation error of the proposed estimator. A debiasing procedure is further employed to yield valid inference. Extensive numerical experiments show that our method can rival or outperform existing approaches in the literature. We finally analyze the mouse obesity data and discuss new findings from our method.

(4) Wei Qian, University of Delaware
Title: *Adaptive Algorithm for Multi-Armed Bandit Problem with High-Dimensional Covariates*

Abstract: This work studies an important sequential decision making problem known as the multi-armed bandit problem with covariates. Under a linear bandit framework with high-dimensional covariates, we propose a general multi-stage arm allocation algorithm that integrates both arm elimination and randomized assignment strategies. By employing a class of high-dimensional regression methods for coefficient estimation, the proposed algorithm is shown to have near optimal finite-time regret performance under a new study scope that requires neither a margin condition nor a reward gap condition for competitive arms. Based on the synergistically verified benefit of the margin, our algorithm exhibits adaptive performance that automatically adapts to the margin and gap conditions, and attains optimal regret rates simultaneously for both study scopes, without or with the margin, up to a logarithmic factor. Besides the desirable regret performance, the proposed algorithm simultaneously generates useful coefficient

estimation output for competitive arms and is shown to achieve both estimation consistency and variable selection consistency. Promising empirical performance is demonstrated through extensive simulation and two real data evaluation examples.

**Session** 41: ***Statistical Application of Functional Data Analysis***
Organizer and Chair: Haolun Shi, Simon Fraser University
**Room: MB252, Time: 1:20 PM – 3:00 PM**

(1) Peijun Sang, University of Waterloo
Title: *Statistical Inference for Functional Linear Quantile Regression*
Abstract: We propose inferential tools for functional linear quantile regression where the conditional quantile of a scalar response is assumed to be a linear functional of a functional covariate. In contrast to conventional approaches, we employ kernel convolution to smooth the original loss function. The coefficient function is estimated under a reproducing kernel Hilbert space framework. A gradient descent algorithm is designed to minimize the smoothed loss function with a roughness penalty. With the aid of the Banach fixed-point theorem, we show the existence and uniqueness of our proposed estimator as the minimizer of the regularized loss function in an appropriate Hilbert space. Furthermore, we establish the convergence rate as well as the weak convergence of our estimator. As far as we know, this is the first weak convergence result for a functional quantile regression model. Pointwise confidence intervals and a simultaneous confidence band for the true coefficient function are then developed based on these theoretical properties. Numerical studies including both simulations and a data application are conducted to investigate the performance of our estimator and inference tools in finite sample. This is a joint work with my collaborators Zuofeng Shang and Pang Du.

(2) Tianyu Guan, Brock University
Title: *Exploring Pre-Launch Movie Electronic Word of Mouth Time Series by Functional Data Analysis*
Abstract: Online product reviews, commonly conceptualized as electronic Word of Mouth, are essentially a time series of multinomial distributions with two dimensions of interest, namely the time dimension and the rating dimension. In this research, we apply functional data analysis to study the data stream of online product reviews so as to find the most efficient way to summarize online product reviews in both the time and rating dimensions in predicting subsequent sales. We observe that most online product review ratings exhibit a positivity bias and extremity, therefore, we apply the functional principal component analysis to explore the major variations among the quantile curves of the movies. The functional principal component (FPC) scores at various quantile levels are then used to predict the box office revenues in the opening week. We use the sparse group LASSO method to select the quantile levels at which the FPC scores make significant contributions to the prediction. In addition, this research shows that top-end percentiles would be better summary statistics compared to the mean

in capturing the relations between the pre-launch product ratings time pattern and launch sales.

(3) Zhiyang Zhou, University of Manitoba
Title: *Smooth Nonparametric Dynamic Prediction for Competing Risks via Deep Learning*
Abstract: In the risk prediction for medicine, public health, economics, engineering, and many other areas, one may have to handle competing risks (i.e., mutually exclusive events) and figure out the relationship between their incidence probabilities and risk factors. Although the recent success of risk prediction has already been extended to the dynamic version (where time-varying risk factors are incorporated into models), existing approaches usually involve strong assumptions (e.g., additive effects and/or proportional hazard) which may lead to extra bias in prediction. To tackle these issues, we present an output layer named the Smooth Monotonic Output Layer (SMOL). When concatenated to deep neural networks, SMOL may help us learn incidence probabilities directly without specifying a parametric structure. We conducted numerical experiments on data collected through the Lifetime Risk Pooling Project (LRPP) which pooled together twenty community-based studies on cardiovascular disease — the leading cause of death in the world — and involved around three hundred thousand participants with long-term follow-ups of longitudinal risk factors. Extensive results showed a state-of-the-art accuracy of our proposal in predicting individual risks of cardiovascular diseases and non-cardiovascular death simultaneously.

(4) Haolun Shi, Simon Fraser University
Title: *A Robust Approach to Functional Principal Component Analysis*
Abstract: It is of great interest to conduct robust functional principal component analysis (FPCA) that can identify the major modes of variation in the stochastic process with the presence of outliers. A new robust FPCA method is proposed in a new regression framework. An M-estimator for the functional principal components is developed based on Huber's loss by iteratively fitting the residuals from the Karhunen-Loveve expansion for the stochastic process under the robust regression framework. Our method can naturally accommodate sparse and irregularly-sampled data. When the functional data have outliers, our method is shown to render stable and robust estimates of the functional principal components; When the functional data have no outliers, we show via simulation studies that the performance of our approach is similar to that of the conventional FPCA method. The proposed robust FPCA method is demonstrated by analyzing some real data sets.

**Coffee Break, MB Central Foyer**

# Parallel Sessions L
# 3:20pm - 5:00pm, Sunday, July 10th

**Session** 42: ***Recent Advances in Causal Identification***

Organizer and Chair: Linbo Wang, University of Toronto
**Room: MB Auditorium, Time: 3:20 PM – 5:00 PM**

(1) Andrew Ying,
Title: *Proximal Causal Inference for Complex Longitudinal Studies*
Abstract: A standard assumption for causal inference about the joint effects of time-varying treatment is that one has measured sufficient covariates to ensure that within covariate strata, subjects are exchangeable across observed treatment values, also known as "sequential randomization assumption (SRA)". SRA is often criticized as it requires one to accurately measure all confounders. Realistically, measured covariates can rarely capture all confounders with certainty. Often covariate measurements are at best proxies of confounders, thus invalidating inferences under SRA. In this paper, we extend the proximal causal inference (PCI) framework of Miao et al. (2018) to the longitudinal setting under a semiparametric marginal structural mean model (MSMM). PCI offers an opportunity to learn about joint causal effects in settings where SRA based on measured time-varying covariates fails, by formally accounting for the covariate measurements as imperfect proxies of underlying confounding mechanisms. We establish nonparametric identification with a pair of time-varying proxies and provide a corresponding characterization of regular and asymptotically linear estimators of the parameter indexing the MSMM, including a rich class of doubly robust estimators, and establish the corresponding semiparametric efficiency bound for the MSMM. Extensive simulation studies and a data application illustrate the finite sample behavior of proposed methods.

(2) Dingke Tang, University of Toronto
Title: *The Synthetic Instrument Method*
Abstract: Inferring causal relationships from observational studies is a predominant problem in social, economical and biomedical sciences. Previous causal studies hinge on either observed confounders or auxiliary variables such as negative controls and instrumental variables. The unconfoundedness condition or the availability of auxiliary variables are often difficult to ensure in modern causal applications. In this paper, we introduce a novel framework that leverages the information contained in multiple causes. Under the commonly used structural equation model and sparsity conditions, we achieve the identification of causal parameters. Furthermore, we develop a simple estimation procedure based on a regularization problem with a specifically designed penalty. We illustrate our framework using breast cancer gene expression data from The Cancer Genome Atlas (TCGA).

(3) Ilya Shpitser, Johns Hopkins University
Title: *The Proximal ID Algorithm*
Abstract: Unobserved confounding is a fundamental obstacle to establishing valid causal conclusions from observational data. Two complementary types of approaches have been developed to address this obstacle: obtaining identification using fortuitous external aids, such as instrumental variables or proxies, or by

means of the ID algorithm, using Markov restrictions on the full data distribution encoded in graphical causal models. In this paper we aim to develop a synthesis of the former and latter approaches to identification in causal inference to yield the most general identification algorithm in multivariate systems currently known – the proximal ID algorithm. In addition to being able to obtain non-parametric identification in all cases where the ID algorithm succeeds, our approach allows us to systematically exploit proxies to adjust for the presence of unobserved confounders that would have otherwise prevented identification. In addition, we outline a class of estimation strategies for causal parameters identified by our method in an important special case. We illustration our approach by simulation studies, and a data application.

(4) Dominik Rothenhaeusler, Stanford University
Title: *Causal Aggregation: Estimation and Inference of Causal Effects by Aggregating Information Across Data Sets*

Abstract: Randomized experiments are the gold standard for causal inference. In experiments, usually, one variable is manipulated and its effect is measured on an outcome. However, practitioners may also be interested in the effect of simultaneous interventions on multiple covariates on a fixed target variable. We discuss a method that allows estimating the effect of joint interventions using data from different experiments in which only very few variables are manipulated. If the joint causal effect is linear, the proposed method can be used for estimation and inference of joint causal effects. We describe extensions to the non-linear and high-dimensional case and discuss conditions under which direct causal effects can be consistently estimated. The proposed method allows to combine data sets arising from randomized experiments as well observational data sets for which IV assumptions or unconfoundedness hold. We demonstrate the effectiveness of the proposed method on synthetic and semi-synthetic data.

**Session** 43: ***High Dimensional Data Analysis***
Organizer and Chair: Xiaoping Shi, The University of British Columbia - Okanagan
**Room: MB253, Time: 3:20 PM – 5:00 PM**

(1) Elham Jamali, University of Calgary
Title: *Doubly Sparse Cox Proportional Hazards Model with a Graphical Structure among Predictors*

Abstract: Abstract By considering predictors as nodes and their correlation as edges, a graphical structure can be imposed on predictors in a survival model. Some researchers have shown that by incorporating the graphical structure into regularized regression, the accuracy of variable selection and inference would increase. In this study, we present a double penalty in penalized likelihood function for the Cox proportional hazards model with a graphical structure among predictors. The proposed double penalty encourages sparsity not only among groups but also at individual levels. Theoretical properties such as the error bound and asymptotic distribution of the regularized estimators are investigated. Our simulation studies show that the proposed method outperforms the existing methods

for the sparse Cox regression model with a graphical structure among predictors. The results of analyzing the pbcseq data also demonstrate the effectiveness of the proposed method.

(2) Augustine Wong, York University
Title: *A Higher Order Likelihood-Based Statistical Inference Procedure for a Vector Parameter of Interest*

Abstract: Statistical inference is a process of using an observed data set to infer properties of a population. This generally resulted in reporting confidence regions of the parameter of interest or reporting the p-values of a significance test. The exact statistical inferential methods are available only for some specific problems. As a result, asymptotic inferential methods are commonly applied. The frequently used asymptotic methods, Wald, Rao and Wilks methods, are based on the likelihood function, and they have only first order accuracy. In recent years, a few likelihood-based third order inferential procedures have been developed; but they are generally restricted to a scalar parameter of interest. In this presentation, a Bartlett-type correction of the Wilks method is proposed. Simulation results show that the proposed method gives extremely accurate coverage even when the sample size is small.

(3) W. John Braun, University of British Columbia
Title: *The Perfect Fire - Segmenting a Video Recording of a Microfire*

Abstract: Video data of a collection of small fires burned in a fume hood are analyzed with a goal to identify regions that are currently burning, burned out and unburned.

(4) Xiaoping Shi, The University of British Columbia - Okanagan
Title: *Two Edge-Count Tests and Relevance Analysis in k High-Dimensional Samples*

Abstract: For the task of relevance analysis, the conventional Tukey's test may be applied to the set of all pairwise comparisons. However, there were few studies that discuss both nonparametric k-sample comparisons and relevance analysis in high dimensions. Our aim is to capture the degree of relevance between combined samples and provide additional insights and advantages in high-dimensional k-sample comparisons. Our solution is to extend a graph-based two-sample comparison and investigate its availability for large and unequal sample sizes. We propose two distribution-free test statistics based on between-sample edge counts and measure the degree of relevance by standardized counts. The asymptotic permutation null distributions of the proposed statistics are derived, and the power gain is proved when the sample sizes are smaller than the square root of the dimension. We also discuss different edge costs in the graph to compare the parameters of the distributions. Simulation comparisons and real data analysis of tumors and images further convince the value of our proposed method. Software implementing the relevance analysis is available in the R package Relevance.

**Session** 44: ***Statistical Disclosure Control Methods for Privacy***
Organizer: Bei Jiang, University of Alberta

Chair: Haihan Xie, University of Alberta
**Room: MB251, Time: 3:20 PM – 5:00 PM**

(1) Yi Liu, University of Alberta
Title: *A Bridge to Gaussian Differential Privacy*
Abstract: Gaussian differential privacy (GDP) is a single-parameter family of privacy notions that provides coherent guarantees to avoid the exposure of sensitive individual information. Relative to DP, GDP provides more interpretability and tighter bounds under composition. Many widely used mechanisms (e.g., the Laplace mechanism) inherently provide GDP guarantees but often fail to take advantage of this new framework because their privacy guarantees were derived under a different background. We develop an easy-to-verify criterion to identify such algorithms and give an efficient method to narrow down possible values of an optimal privacy measurement, $\mu$ with an arbitrarily small and quantifiable margin of error.

(2) Bei Jiang, University of Alberta
Title: *Creation of Privacy-Preserving Synthetic Data for Research Reproducibility, with Application to Patient Registry Data*
Abstract: Responsible data sharing anchors research reproducibility, which in turn promotes integrity of scientific research. However, due to privacy and confidentiality concerns, restricted-use research data has been difficult to access. Motivated by the Canadian Scleroderma Research Group (CSRG) patient registry data, we present a risk-based masking approach to produce privacy-preserving synthetic datasets, which also simultaneously imputes the missing data of mixed continuous and categorical types in the original dataset. In contrast to a one-size-fits-all strategy, this approach divides all individuals into different subgroups based on their re-identification risks, and provides masking strategies targeted for each risk subgroup, through the associated tuning mechanism. This risk-based approach reduced the number of patients at risk of re-identification from 193 to 7, out of the 698 CSRG patients, while preserving all correct inferential conclusions in the target analysis, with the 95% confidence intervals (CIs) having 95.6% overlaps on average with the CIs constructed using the full and unmasked datasets. Other competing approach led to lower CI-overlaps (on average 61.6%), ranging from -118.4% (no overlap) to 97%, and as a result some incorrect inferential conclusions. These findings suggest that our risk-based masking approach makes it possible to release full synthetic datasets for research reproducibility purposes while ensuring that the re-identification risks are acceptably low.

(3) Fang Liu, University of Notre Dame
Title: *A New Bound for Privacy Loss from Bayesian Posterior Sampling*
Abstract: Differential privacy (DP) is a state-of-the-art concept that formalizes privacy guarantees. We derive a new bound for the privacy loss from releasing Bayesian posterior samples in the setting of DP. The new bound is tighter than the existing bounds for common Bayesian models and is also consistent with the likelihood principle. We apply the privacy loss quantified by the new bound

to release differentially private synthetic data from Bayesian models in several experiments and show the improved utility of the synthetic data compared to those generated from explicitly designed randomization mechanisms that privatize posterior distributions.

(4) Wei Tu, Queen's University
Title: *Differential Privacy with Survival Data*
Abstract: The protection of individual patient privacy is essential in health care research. Privacy-protecting data analysis has a long history under the name of "statistical disclosure control" in statistics. Differential privacy, emerging from the theoretical computer science literature, has become popular over the last decade due to its intuitive formulation and formal privacy guarantee, and is at its early stages of implementation in industry, government and academia. In this talk, I will present the framework of differential privacy and present a few applications in health research. Specifically, a differentially private Kaplan-Meier estimate using the recently proposed Gaussian differential privacy framework will be presented, as well as differential private learning in training clinical prediction task using EHR and medical imaging data.

**Session** 45: ***Advances in the Analysis of Complex Lifetime Data***
Organizer and Chair: Hua Shen, University of Calgary
**Room: MB252, Time: 3:20 PM – 5:00 PM**

(1) Renjun Ma, University of New Brunswick
Title: *Survival Analysis of Car Accident Data While Accounting for Partially Crossed Location and Agent Effects*
Abstract: In automobile insurance studies, car accident data are often partially cross cross-classified by location and agent. One research question of great interest is to link time to occurrence of car accident with various factors. An appropriate analysis of such data needs to account for location and agent effects. In this talk, we incorporate partially crossed random effects into Cox proportional hazards models for such data and propose a Poisson modeling approach to model estimation. We predict the random effects using the orthodox best linear unbiased predictor method, and obtain consistent estimators for the regression parameters. This estimating method relies on only the first and second moments of the random effects. Our approach is illustrated with a collection of large automobile insurance data. Another potential application of our approach is to study clinical data partially cross-classified by residential areas and medical service providers. This is joint work with Shi Zhang and Guohua Yan.

(2) Karen Kopciuk, University of Calgary
Title: *Generating and Modelling Time-to-Event Data for Family Study Designs in Genetics Applications*
Abstract: Family-based designs are popular in genetic studies for several reasons, including robustness to population admixture and inclusion of more individuals who are gene carriers. However, the clustering of related individuals, missing

genetic information and family selection are issues that need to be addressed in analyses of data obtained from these designs. In this talk I will provide an overview of family-based studies, then describe the analyses of time-to-event data that estimates the age at onset of disease correcting for common sources of bias and sampling of families. Simulation study options as well as application to sample size and power calculations for planning a study will also be presented. This work is based on our user-friendly R package FamEvent.

(3) Yildiz Yilmaz, Memorial University
Title: *Multi-State Cure Modeling of Cancer Progression*

Abstract: Analyses of disease-free survival data for some cancer types indicate that cohorts of patients treated for cancer consist of individuals who are susceptible to experience cancer related events and individuals who are cured. Cured individuals do not experience any cancer related event, and eventually die due to other causes. Individuals who are not cured may die after experiencing a cancer recurrence or without experiencing any recurrence. Cure status is a partially latent variable and is only known if a disease related event is observed. To model disease progression events, we consider a semi-Markov multi-state model including partially latent cured and not cured states. In this talk, I will describe our modeling approach and discuss an inference method handling masked causes of deaths in addition to partially latent cure status. This is a joint study with Yongho Lim and Candemir Cigsar.

(4) Hua Shen, University of Calgary
Title: *Analysis of Recurrent Events with a Misclassified Covariate*

Abstract: In the analysis of recurrent events data, standard methods require the covariates to be completely and precisely observed. However, misclassification in covariate often arise and naïve use of misclassified covariate results in biased estimator of covariate effects on the risk of event occurrence. We develop a likelihood-based method to fit regression models to recurrent event data involving misclassified covariate in the absence of validation data. The likelihood-based algorithm is shown to yield estimators with small empirical bias and much advantageous than the naive approaches in simulation studies.

**Session** 46: ***Statistical Learning Methods and Applications***
Organizer and Chair: Kaiqiong Zhao, University of Alberta
**Room: Elder Tom Crane Bear, Time: 3:20 PM – 5:00 PM**

(1) Qiongshi Lu, University of Wisconsin-Madison
Title: *Benchmarking and Fine-Tuning Prediction Models with Marginal Summary Statistics*

Abstract: Polygenic risk scores (PRSs) trained from marginal summary statistics of genome-wide association studies (GWAS) have broad applications in human genetics research. However, due to the highly summarized nature of the input data, statistical learning tasks that rely on cross-validation (e.g. selecting tuning parameters) cannot be performed. Here, we introduce PUMAS (Parameter-

tuning Using Marginal Association Statistics), a novel method to perform cross-validation using summary statistics from GWAS. PUMAS has two key steps. First, we sample marginal association statistics for a subset of individuals based on the complete GWAS summary statistics to generate both training and testing data. Second, we evaluate the predictive performance of PRS using the resampled summary statistics. In this talk, I will also introduce recent extensions of PUMAS which allow the framework to optimize penalized regression models and combine multiple scores through regression modeling using only summary data as input. I will demonstrate the performance of PUMAS through benchmarking and optimizing PRS performance for hundreds of complex traits with publicly available GWAS summary statistics. We believe our method resolves a fundamental problem without a current solution and will greatly benefit genetic prediction applications.

(2) Quefeng Li, University of North Carolina at Chapel Hill
Title: *Integrative Factor Regression and Its Inference for Multimodal Data Analysis*

Abstract: Multimodal data, where different types of data are collected from the same sub- jects, are fast emerging in a large variety of scientific applications. Factor analysis is commonly used in integrative analysis of multimodal data, and is particularly useful to overcome the curse of high dimensionality and high correlations. However, there is little work on statistical inference for factor analysis based supervised modeling of multimodal data. In this article, we consider an integrative linear regression model that is built upon the latent factors extracted from multimodal data. We address three important questions: how to infer the significance of one data modality given the other modalities in the model; how to infer the significance of a combination of variables from one modality or across different modalities; and how to quantify the contribution, measured by the goodness-of-fit, of one data modality given the others. When answering each question, we explicitly characterize both the benefit and the extra cost of factor analysis. Those questions, to our knowledge, have not yet been addressed despite wide use of factor analysis in integrative multimodal analysis, and our proposal bridges an important gap. We study the empirical performance of our methods through simulations, and further illustrate with a multimodal neuroimaging analysis.

(3) Lucy Gao, University of Waterloo
Title: *Inference after Latent Variable Estimation in Single-Cell RNA-Sequencing Data*

Abstract: In the context of single-cell RNA-sequencing data, we often wish to perform unsupervised learning of latent structure among the cells, and then test for association between this latent structure and gene expression. For example, we might estimate cell types via clustering, and then test whether gene expression differs across cell types. Alternatively, we might estimate a low-dimensional subspace representing a continuous cellular developmental trajectory, and then test for association between gene expression and this trajectory. Unfortunately,

classical tests of association between gene expression and the latent structure will not control the type I error rate, since the same data is used to estimate the latent structure and to perform hypothesis testing. Furthermore, sample-splitting - i.e. splitting the cells into independent train and test sets - does not solve the problem.

In this talk, I will discuss two solutions to this problem. The first involves an application of selective inference, and the second involves "count splitting", a simple alternative way to split the data into independent train and test sets that does control the type I error rate.

# List of Participants

Note: late registrants are not listed here.

| Last Name | First Name | Organization | Email |
|---|---|---|---|
| Agustin | Mayo-Iscar | Universidad De Valladolid | agustin.mayo.iscar@uva.es |
| Asif | Neloy | University of Manitoba | neloyn@myumanitoba.ca |
| Augustine | Wong | York University | august@yorku.ca |
| Bei | Jiang | University of Alberta | bei1@ualberta.ca |
| Benjamin | Risk | Emory University | benjamin.risk@emory.edu |
| Bin | Li | University of Waterloo | bin.li@uwaterloo.ca |
| Bing | Li | Penn State University | bxl9@psu.edu |
| Ce | Zhang | University of Alberta | ce5@ualberta.ca |
| Ce | Zhang | University of Alberta | cezhang0321@gmail.com |
| Chad | He | Fred Hutchinson Cancer Research Center | qhe@fredhutch.org |
| Changgee | Chang | University of Pennsylvania | changgee@pennmedicine.upenn.edu |
| Danika | Lipman | University of Calgary | danilipman@gmail.com |
| Dehan | Kong | University of Toronto | dehan.kong@utoronto.ca |
| Dengdeng | Yu | University of Texas at Arlington | dengdeng.yu@uta.edu |
| Depeng | Jiang | University of Manitoba | depeng.jiang@umanitoba.ca |
| Dianliang | Deng | University of Regina | deng@uregina.ca |
| Dingke | Tang | University of Toronto | dingke.tang@mail.utoronto.ca |
| Dominik | Rothenhaeusler | Stanford University | rdominik@stanford.edu |
| Duncan | Fong | Pennsylvania State University | i2v@psu.edu |
| Eardi | Lila | University of Washington | elila@uw.edu |
| Elham | Jamali | University of Calgary | elham.jamali@ucalgary.ca |
| Emily | Hector | North Carolina State University | ehector@ncsu.edu |
| Enze | Shi | University of Alberta | eshi@ualberta.ca |
| Esra | Kurum | University of California, Riverside | esra.kurum@ucr.edu |
| Faming | Liang | Purdue University | fmliang@purdue.edu |
| Fan | Yang | University of Waterloo | fan.yang@uwaterloo.ca |
| Fang | Han | University of Washington | fanghan@uw.edu |
| Fang | Liu | University of Notre Dame | fliu2@nd.edu |
| Farouk | Nathoo | University of Victoria | nathoo@uvic.ca |
| Fei | Gao | Fred Hutchinson Cancer Center | fgao@fredhutch.org |
| Gen | Li | University of Michigan | ligen@umich.edu |
| Guohua | Yan | University of New Brunswick | gyan@unb.ca |
| Hai | Shu | Nyu School of Global Public Health | hs120@nyu.edu |
| Haihan | Xie | University of Alberta | haihan1@ualberta.ca |
| Haiying | Wang | University of Connecticut | haiying.wang@uconn.edu |
| Hao | Chen | University of California, Davis | hxchen@ucdavis.edu |
| Haolun | Shi | Simon Fraser University | haoluns@sfu.ca |
| Heping | Zhang | Yale University | heping.zhang@yale.edu |
| Hong | Gu | Dalhousie University | hgu@dal.ca |
| Hongtu | Zhu | The University of North Carolina at Chapel Hill | htzhu@email.unc.edu |
| Hongzhe | Li | University of Pennsylvania | hongzhe@pennmedicine.upenn.edu |
| Hua | Shen | University of Calgary | hua.shen@ucalgary.ca |
| Hua | Zhou | University of California, Los Angeles | huazhou@ucla.edu |
| Hui | Zhang | Northwestern University | hzhang@northwestern.edu |
| Huixia | Wang | National Science Foundation | huiwang@nsf.gov |
| Ilya | Shpitser | Johns Hopkins University | ilyas@cs.jhu.edu |
| Ivan | Mizera | University of Alberta | imizera@ualberta.ca |
| Janie | Coulombe | Mcgill University | janie.coulombe@mail.mcgill.ca |
| Jessica | Gronsbell | University of Toronto | j.gronsbell@utoronto.ca |
| Jian | Kang | University of Michigan | jiankang@umicu.edu |
| Jianqing | Fan | Princeton University | jqfan@princeton.edu |
| Jiaying | Gu | University of Toronto | jiaying.gu@utoronto.ca |
| Jin | Zhou | University of California, Los Angeles | jin.jjzhou@gmail.com |
| Jinchi | Lv | University of Southern California | jinchilv@marshall.usc.edu |
| Jinhan | Xie | University of Alberta | jinhan3@ualberta.ca |
| Joan | Hu | Simon Fraser University | joanh@stat.sfu.ca |
| Joy | Jiang | Washington University School of Medicine | jiang.shu@wustl.edu |
| Julien | St-Pierre | Mcgill University | stpierre.ju@gmail.com |
| Junhao | Zhu | University of Toronto | jh.zhu@mail.utoronto.ca |
| Juxin | Liu | University of Saskatchewan | liu@math.usask.ca |
| Kai | Wang | University of Iowa | kai-wang@iowa.edu |
| Kaiqiong | Zhao | University of Alberta | kaiqiong@ualberta.ca |
| Karen | Kopciuk | University of Calgary | kakopciu@ucalgary.ca |

| | | | |
|---|---|---|---|
| Katarzyna | Reluga | University of Toronto | katarzynareluga@gmail.com |
| Ke | Sun | University of Alberta | ksun6@ualberta.ca |
| Kevin | Mcgregor | York University | kevinmcg@yorku.ca |
| Kevin | Zhang | University of Toronto | kevinkw.zhang@mail.utoronto.ca |
| Kuan | Liu | University of Toronto | kuan.liu@utoronto.ca |
| Kwun Chuen Gary | Chan | University of Washington | kcgchan@uw.edu |
| Lan | Luo | The University of Iowa | lan-luo@uiowa.edu |
| Lan | Xue | Oregon State University | xuel@stat.oregonstate.edu |
| Lang | Wu | University of British Columbia | lang@stat.ubc.ca |
| Laura | Cowen | University of Victoria | lcowen@uvic.ca |
| Lei | Ding | University of Alberta | lding1@ualberta.ca |
| Lei | Sun | University of Toronto | sun@utstat.toronto.edu |
| Leilei | Zeng | University of Waterloo | lzeng@uwaterloo.ca |
| Li | Xing | University of Saskatchewan | lix491@usask.ca |
| Liangliang | Wang | Simon Fraser University | lwa68@sfu.ca |
| Liangyuan | Hu | Rutgers University | lh707@sph.rutgers.edu |
| Lily | Wang | George Mason University | lwang41@gmu.edu |
| Linbo | Wang | University of Toronto | linbo.wang@utoronto.ca |
| Linglong | Kong | University of Alberta | lkong@ualberta.ca |
| Lingsong | Zhang | Purdue University | lingsong@purdue.edu |
| Lingzhu | Li | University of Alberta | lingzhu@ualberta.ca |
| Liqun | Diao | University of Waterloo | l2diao@uwaterloo.ca |
| Liqun | Wang | University of Manitoba | liqun.wang@umanitoba.ca |
| Longhai | Li | University of Saskatchewan | longhai.li@gmail.com |
| Luca | Bagnato | Università Cattolica Del Sacro Cuore | luca.bagnato@unicatt.it |
| Lucy | Gao | University of Waterloo | lucy.gao@uwaterloo.ca |
| Madeline | Ward | University of Calgary | madeline.ward1@ucalgary.ca |
| Matias | Salibian Barrera | The University of British Columbia | matias@stat.ubc.ca |
| Matus | Maciak | Charles Univerzity | matus.maciak@mff.cuni.cz |
| Maxime | Turgeon | University of Manitoba | max.turgeon@umanitoba.ca |
| Md | Mahsin | University of Calgary | md.mahsin@ucalgary.ca |
| Mei | Li | University of Alberta | mei7@ualberta.ca |
| Meichen | Liu | University of Alberta | meichen1@ualberta.ca |
| Michael | Wu | Fred Hutchinson Cancer Center | mcwu2004@gmail.com |
| Michal | Pesta | Charles Univerzity | michal.pesta@mff.cuni.cz |
| Mingqi | Wu | McGill University | mingqi.wu@mail.mcgill.ca |
| Mireille | Schnitzer | Université De Montréal | mireille.schnitzer@umontreal.ca |
| Mohammad Ehsanul | Karim | The University of British Columbia | ehsan.karim@ubc.ca |
| Mosuk | Chow | Statistics Department, Penn State University | mchow@stat.psu.edu |
| Mu | Zhu | University of Waterloo | m3zhu@uwaterloo.ca |
| Na | Zhang | University of Alberta | nz4@ualberta.ca |
| Nanwei | Wang | University of New Brunswick | nanwei.wang@unb.ca |
| Nicola | Gnecco | University of Geneva | nicola.gnecco@unige.ch |
| Ning | Hao | University of Arizona | nhao@math.arizona.edu |
| Pan | Bo | University of Alberta | pan1@ualberta.ca |
| Pang | Du | Virginia Tech | pangdu@vt.edu |
| Peijun | Sang | University of Waterloo | psang@uwaterloo.ca |
| Peisong | Han | University of Michigan | peisong@umich.edu |
| Peter | Song | University of Michigan | pxsong@umich.edu |
| Qingxia | Chen | Vanderbilt University Medical Center | cindy.chen@vumc.org |
| Qiongshi | Lu | University of Wisconsin-Madison | qlu@biostat.wisc.edu |
| Quefeng | Li | University of North Carolina at Chapel Hill | quefeng@email.unc.edu |
| Radu | Craiu | University of Toronto | radu.craiu@utoronto.ca |
| Renjun | Ma | University of New Brunswick | renjun@unb.ca |
| Robert | Ogden | Columbia University | to166@columbia.edu |
| Ruoqing | Zhu | University of Illinois Urbana Champaign | rqzhu@illinois.edu |
| Sahir | Bhatnagar | Mcgill University | sahir.bhatnagar@mcgill.ca |
| Sebastian | Engelke | University of Geneva | sebastian.engelke@unige.ch |
| Shu | Yang | North Carolina State University | syang24@ncsu.edu |
| Shujie | Ma | University of California, Riverside | shujie.ma@ucr.edu |
| Stanislav | Volgushev | University of Toronto | stanislav.volgushev@utoronto.ca |
| Teng | Zhang | University of Central Florida | teng.zhang@ucf.edu |
| Thierry | Chekouo | University of Calgary | thierry.chekouotekou@ucalgary.ca |
| Thorsten | Koch | Tu Berlin | koch@zib.de |
| Tianyu | Guan | Brock University | tguan@brocku.ca |
| Tingting | Zhang | University of Pittsburgh | tiz67@pitt.edu |
| Toby | Kenney | Dalhousie University | tkenney@mathstat.dal.ca |
| Wei | Qian | University of Delaware | weiqian@udel.edu |

| | | | |
|---|---|---|---|
| Wei | Sun | Purdue University | sun244@purdue.edu |
| Wei | Tu | Queen's University | wei.tu@queensu.ca |
| Weibin | Mo | Purdue Unviersity | harrymok@email.unc.edu |
| Wen | Zhou | Colorado State University | riczw@stat.colostate.edu |
| Wendy | Lou | University of Toronto | wendy.lou@utoronto.ca |
| Wenxin | Zhou | Uc San Diego | wez243@ucsd.edu |
| Willard | Braun | University of British Columbia | john.braun@ubc.ca |
| Wilson | Lu | Acadia University | wen.wilsonlu@gmail.com |
| Xiao | Wang | Purdue University | wangxiao@purdue.edu |
| Xiaodong | Yan | Shandong University | yanxiaodong@sdu.edu.cn |
| Xiaofeng | Wang | Cleveland Clinic | wangx6@ccf.org |
| Xiaojian | Xu | Brock University | xxu@brocku.ca |
| Xiaoke | Zhang | The George Washington University | xkzhang@gwu.edu |
| Xiaomeng | Ju | University of British Columbia | xiaomeng.ju@stat.ubc.ca |
| Xiaoping | Shi | The University of British Columbia - Okanagan | xiaoping.shi@ubc.ca |
| Xiaotong | Shen | University of Minnesota | xshen@umn.edu |
| Xiaowu | Dai | University of California, Berkeley | xwdai@berkeley.edu |
| Xikui | Wang | University of Manitoba | xikui.wang@umanitoba.ca |
| Xinping | Cui | University of California, Riverside | xinping.cui@ucr.edu |
| Xinyi | Zhang | University of Toronto | zhangxinyimars2@gmail.com |
| Xuekui | Zhang | University of Victoria | xuekui@uvic.ca |
| Xuewen | Lu | University of Calgary | xlu@ucalgary.ca |
| Yafei | Wang | University of Alberta | yafei2@ualberta.ca |
| Yan | Cui | University of Toronto | yyan.cui@mail.utoronto.ca |
| Yan Shuo | Tan | University of California, Berkeley | yanshuo@gmail.com |
| Yanyuan | Ma | Pennsylvania State University | yanyuanma@yahoo.com |
| Yao | Luo | University of Toronto | yao.luo@utoronto.ca |
| Yehua | Li | University of California, Riverside | yehuali@ucr.edu |
| Yeying | Zhu | University of Waterloo | yeying.zhu@uwaterloo.ca |
| Yi | Lian | Mcgill University | yi.lian@mail.mcgill.ca |
| Yi | Liu | University of Alberta | yliu16@ualberta.ca |
| Yi | Xiong | Fred Hutchinson Cancer Research Center | yxiong@fredhutch.org |
| Yi | Yang | Mcgill University | archer.yang@mcgill.ca |
| Yi | Zhao | Indiana University | yz125@iu.edu |
| Yichi | Zhang | North Carolina State University | yzhan239@ncsu.edu |
| Yildiz | Yilmaz | Memorial University | yyilmaz@mun.ca |
| Ying | Chen | National University of Singapore | matcheny@nus.edu.sg |
| Ying | Chen | National University of Singapore | matcheny@nus.edu.sg |
| Ying | Zhang | Acadia University | ying.zhang@acadiau.ca |
| Ying | Zhou | University of Toronto | yingx.zhou@mail.utoronto.ca |
| Yingwei | Peng | Queen's University | pywpeng@gmail.com |
| Yingying | Fan | University of Southern California | fanyingy@usc.edu |
| Yue | Niu | University of Arizona | yueniu@math.arizona.edu |
| Yuying | Xie | Michigan State University | xyy@msu.edu |
| Yuzi | Liu | University of Alberta | yuzi3@ualberta.ca |
| Zehui | Wang | Queen's University | 20zw8@queensu.ca |
| Zhao | Ren | University of Pittsburgh | zren@pitt.edu |
| Zhaohan | Sun | University of Waterloo | z227sun@uwaterloo.ca |
| Zheng | Yu | University of Calgary | zheng.yu1@ucalgary.ca |
| Zhengwu | Zhang | University of North Carolina at Chapel Hill | zz10c@email.unc.edu |
| Zhenhua | Lin | National University of Singapore | linulysses@gmail.com |
| Zhenhua | Lin | National University of Singapore | linz@nus.edu.sg |
| Zhezhen | Jin | Columbia University | zj7@cumc.columbia.edu |
| Zhiwen | Tan | Queen's University | 21zt9@queensu.ca |
| Zhixian | Yang | University of Alberta | zhixian@ualberta.ca |
| Zhiyang | Zhou | University of Manitoba | zhiyang.zhou@umanitoba.ca |
| Zhiyu | Quan | University of Illinois at Urbana-Champaign | zquan@illinois.edu |
| Ziang | Niu | University of Pennsylvania | ziangniu@sas.upenn.edu |
| Zuofeng | Shang | New Jersey Institute of Technology | zshang@njit.edu |