Welcome!

Statistics Graduate Student Research Day

April 27, 2023

Statistical Sciences
UNIVERSITY OF TORONTO

About

Statistics Graduate Student Research Day

Statistics Graduate Student Research Day is an annual research conference organized by graduate students in statistical sciences at the University of Toronto. The purpose of this conference is to highlight exciting work by graduate students in the areas of statistics, mathematical science, and actuarial science. The conference also serves as a networking event, bringing together graduate students from other departments and/or universities with our own statistical sciences community.

Useful Information

Statistics Graduate Student Research Day will be taking place in the Department of Statistical Sciences at the University of Toronto. The address of the department is *9th Floor, Ontario Power Building 700 University Avenue*. Talks will be held in the Maple and Cherry Rooms (9016 and 9014). The talks will start at 10am, with opening remarks beginning at 9:45am. The conference will end at approximately 17:05, after closing remarks. For a schedule, see the next page. The list of abstracts follows. If you have any questions, please do not hesitate to reach out to a member of the organizing committee.

Acknowledgement of Traditional Land

We wish to acknowledge this land on which the University of Toronto operates. For thousands of years it has been the traditional land of the Huron-Wendat, the Seneca, and the Mississaugas of the Credit. Today, this meeting place is still the home to many Indigenous people from across Turtle Island and we are grateful to have the opportunity to work on this land.

Organizing committee

Sophia Chan
Jianhui Gao
Emma Kroell
Xiochuan Shi
Yuan Tian
Liam Welsh
Ziang Zhang
Ying Zhou

The organizing committee thanks Dr. Monica Alexander and Dr. Stanislav Volgushev for assistance and guidance in planning. The organizing committee also thanks Dr. Michael Evans and the Department of Statistical Sciences at U of T for providing funding support.

Timetable

Affiliations: UoT – University of Toronto; UW – University of Waterloo
Chairs: YJ – Yovna Junglee; LW – Liam Welsh; YZ – Ying Zhou; ZZ – Ziang Zhang

Thursday, April 27

9:45–10:00	**Opening remarks**		
10:00 – 11:00	LW	**Anthony Coache** UoT	Risk-Aware Reinforcement Learning for Dynamic Risk Measures
11:00–12:00	LW	**Luke Hagar** UW	Fast Sample Size Determination for Bayesian Equivalence Tests
12:00–13:00	**Lunch (Provided)**		
13:00–13:30	YZ	**Morris Greenberg** UoT	Restricted Search Space MCMC with Adaptive Weighting and Sparsity Parameterization for Graph Inference
13:30–14:00	YZ	**Vedant Choudhary** UoT	Simulating Implied Volatility Surfaces with Neural SDEs
14:00–14:15	**Break**		
14:15–14:45	YJ	**Alexandra Mossman** UW	Dynamic Treatment Regimes for Clustered and Hierarchical Data with Interference
14:45–15:15	YJ	**Yuan Tian** UoT	Leveraging Multimodal Neuroimaging Data to Identify Novel Genetic Pathways to Alzheimer’s Disease
15:15–15:30	**Break**		
15:30-16:00	ZZ	**Mufan (Bill) Li** UoT	Singular Values for Products of Ginibre Matrices and Neural Networks at Initialization
16:00-16:30	ZZ	**Sonia Markes** UoT	Effect of the Infield Shift: Causal Inference in Baseball
16:30–17:00	ZZ	**Emily Somerset** UoT	Modeling age patterns of migration: a flexible framework to account for unexpected deviations
17:00–17:05	**Closing remarks**		

List of Abstracts

Categories: AS – Applied Statistics; BS – Biostatistis; CS – Computational Statistics; DS – Data Science; MF – Mathematical Finance; ML– Machine Learning; TS – Theoretical Statistics

Risk-Aware Reinforcement Learning for Dynamic Risk Measures

Anthony Coache AS, MF, ML

University of Toronto

Most reinforcement learning (RL) approaches seek to optimize discounted rewards for a risk-neutral agent. While there is some work in risk-aware RL, they typically provide optimal precommitment strategies, are tuned to a specific measure of risk, or are applicable only in small state-action spaces or other simplified settings. Here, instead, we make use of dynamic risk measures to assess the risk of a sequence of costs, and develop a risk-aware RL approach for optimizing strategies. We illustrate the efficacy of our framework within financial applications. Joint work with Sebastian Jaimungal and Álvaro Cartea.

Fast Sample Size Determination for Bayesian Equivalence Tests

Luke Hagar AS, CS

University of Waterloo

Equivalence testing allows one to conclude that two characteristics are practically equivalent. We propose a framework for fast sample size determination with Bayesian equivalence tests facilitated via posterior probabilities. We assume that data are generated using statistical models with fixed parameters for the purposes of sample size determination. Our framework leverages an interval-based approach, which defines a distribution for the sample size to control the length of posterior highest density intervals (HDIs). We prove the normality of the limiting distribution for the sample size, and we consider the relationship between posterior HDI length and the statistical power of Bayesian equivalence tests. We introduce two novel approaches for estimating the distribution for the sample size, both of which are calibrated to align with targets for statistical power. Both approaches are much faster than traditional power calculations for Bayesian equivalence tests. Moreover, our method requires users to make fewer choices than traditional simulation-based methods for Bayesian sample size determination, which makes it more accessible to users accustomed to frequentist methods.

Restricted Search Space MCMC with Adaptive Weighting and Sparsity Parameterization for Graph Inference

Morris Greenberg CS

University of Toronto

Inferring a directed acyclic graph (DAG) given data is computationally challenging due to the large search space needed to exhaustively consider all possible graphs. Current state-of-the-art MCMC methods for graph inference efficiently scan the space by first considering a restricted search space and iteratively expand the space until a stopping criterion is met. Here, we find conditions on iterative changes to the search space that allow for posterior convergence on the unrestricted space, and develop a novel MCMC method that satisfies these conditions. Our algorithm allows for both expansion and constriction of the search space at any iteration, and allows for larger expansion steps than previous methods allow for. Our expansion procedure is dictated by parameterizing the sparsity of the graph (the maximal number of parents observed), and considering how it induces realized parent sizes for individual vertices in the graph, and our constriction procedure is dictated by associating a weight (determined by previous steps in the MCMC chain) to each edge in the space, and only considering edges with large enough weights. We present extensive simulations that characterize the performance and computational efficiency of our algorithm, contrast this with existing methods, and consider applications in the field of imaging proteomics.

Simulating Implied Volatility Surfaces with Neural SDEs

Vedant Choudhary DS, MF, ML

University of Toronto

In this talk, we develop a market simulator for generating sequences of implied volatility surfaces and the corresponding equity prices using a combination of functional data analysis and neural stochastic differential equations (SDEs). We demonstrate that learning the joint dynamics, across multiple assets, of surfaces and prices produces consistent market scenarios across the assets that are consistent with historical features. The sequence of generated surfaces lie within the sub-manifold of surfaces that are free of static arbitrage without the need to explicitly impose such constraints. Moreover, the simulated price paths include those that represent extreme market behaviour, thus allowing users to probe tail risk in their scenarios.

Dynamic Treatment Regimes for Clustered and Hierarchical Data with Interference

Alexandra Mossman AS, BS

University of Waterloo

Dynamic treatment regimes (DTRs) aim to optimize a patient's outcome by using available information at each stage of follow-up to recommend a particular course of treatment. DTRs are often applied towards observational datasets under the stable unit treatment value assumption (SUTVA), which states that an individual's outcome is independent of the treatments received by others at each stage of follow-up. In practice, this assumption is often violated due to different forms of interference that occurs within social networks; for instance, an individual may be less likely to become infected if their neighbours are fully vaccinated, demonstrating an indirect effect of neighbours' treatment in addition to the direct effect of vaccination status of the given individual. Although much of the biostatistical literature to date considers partial interference, where interference occurs within groups but not across groups, much work has yet to be done for modifying DTR methodology to also account for between-group interference in clustered and hierarchical datasets. This work introduces a modification of dynamic weighted ordinary least squares regression (dWOLS) as a DTR method that uses network weights based on propensity score functions modelling both fixed and random effects within groups of patients to account for individual- and group-specific factors affecting the nature of interference present.

Leveraging Multimodal Neuroimaging Data to Identify Novel Genetic Pathways to Alzheimer's Disease

Yuan Tian AS, BS

University of Toronto

Recent genome-wide association studies (GWASs) have identified multiple genetic risk factors for Alzheimer's disease (AD). However, they do not provide a comprehensive understanding of how genetically-regulated structural and functional brain pathways drive AD progression, which is critical for characterizing the genetic mechanism of AD and developing AD targeted therapeutics. We propose a three-step mediation method for analyzing causal pathways of gene-AD effects by incorporating high-dimensional and multimodal brain magnetic resonance imaging (MRI) measures as mediators in GWAS. First, to reduce high dimensionality of MRI while preserving shared structure across brain modalities, we apply BIDIFAC, a dimension reduction method that extends the idea of principal components (PCA). Next, we estimate the genetic variations of reduced-dimensional MRI features using penalized regression. Lastly, we test for the existence of intermediate causal effects between genes and AD with an adaptive association test. We apply the proposed method to UK Biobank (UKB) and International Genomics of Alzheimer's Project (IGAP) data to identify novel genetic pathways to AD.

Singular Values for Products of Ginibre Matrices and Neural Networks at Initialization

Mufan (Bill) Li ML, TS

University of Toronto

It is well known that the spectral distribution for a single symmetric random matrix converges to the semicircle law, where it is closely characterized by the Stieltjes transform and a diffusion process called Dyson Brownian motion. We show that iid. products of Ginibre matrices has a similar structure with a modified version of both Stieltjes transform and Dyson Brownian motion characterizing its singular value distribution. In the shaped limit of neural networks at initialization, we can also approximate the spectral distribution of the covariance kernel matrix.

Effect of the Infield Shift: Causal Inference in Baseball

Sonia Markes AS

University of Toronto

The infield shift is defensive strategy in baseball that has been used with increasing frequency in recent years. Along with the trend in its usage, notoriety of the shift has grown, as it is believed to be responsible for the recent decline in offence. For the 2023 season, Major League Baseball (MLB) will implement a rule change prohibiting the infield shift. However, there has been no systematic analysis of the effectiveness of infield shift to determine if it is a cause of the cooling in offence. We used publicly available data on Major League Baseball from 2015-2021 to evaluate the causal effect of the infield shift on the expected runs scored. We employed three methods for drawing causal conclusions from observational data—nearest neighbour matching, weighting by the odds, and instrumental variable analysis—and looked at subgroups defined by batter handedness. While exact estimates varied, the results of all methods showed the shift is effective at preventing runs, but only for left handed batters.

Modeling age patterns of migration: a flexible framework to account for unexpected deviations

Emily Somerset AS

University of Toronto

We propose a general model framework to estimate age-specific migration rates at the subnational level. The model can be decomposed into an expected level — which consists of an overall mean age schedule plus a local area level — and deviations away from expected, which are smoothed over age and time. This modeling framework allows for reliable estimates of local area migration by age to be made, and also allows deviations away from the expected level to be understood from a temporal perspective. We show results as applied to in-migration rates to puma levels in the United States and show how local-area migration rates have been lower (or higher) than expected since the onset of the Covid-19 pandemic. Future work will extend this framework to county-level migration rates.