Welcome!

Statistics Graduate Student Research Day

Friday April 24, 2026

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 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*, *Intact Center*, *700 University Avenue*, *Toronto*, *ON*, *M5G 1Z5*. The talks and poster session will be held in the Maple and Cherry Rooms (9014 and 9016). A small breakfast (pastries and coffee) will be provided at 9:30am. Opening remarks will begin at 9:45am, followed by the first round of talks at 10am. Lunch will be provided at 12:45pm. The poster session will be held at 2:30pm. The conference will conclude at approximately 5pm, after closing remarks. For a complete schedule, see the next page. The list of abstracts follows after that.

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

Luis Sierra Muntané	Pak Hop Chan	Pat Chimtanoo
Elijah French	William Groff	Bernard Miškić
Noah Ripstein	KC Tsiolis	Vinky Wang
	Mahdi Zamani	

The organizing committee thanks faculty member Dr. Stanislav Volgushev, as well as staff members Karla Barrera, Kachely Peters and Donna Hill for their assistance and guidance in planning. The organizing committee also thanks the Department of Statistical Sciences at the University of Toronto for providing funding support.

Timetable

Friday, April 24, 2026

9:45–10:00	**Opening Remarks**		
	Block 1: Data Science Theory		
10:00 – 10:30	PC	**Bingqing Li** DoSS	Convergence and Optimality of the EM Algorithm Under Multi-Component Gaussian Mixture Models
10:30–11:00	PC	**Luis Sierra Muntané** DoSS	Quantifying Optimism and Model Comparison in Matrix Approximation: a Degrees of Freedom Approach
11:00-11:15	**Break**		
	Block 2: Applied Statistics		
11:15–11:45	BM	**Rodrigo Herrera** DoSS	Scalable Bayesian Additive Models for Stellar Flare Detection via Amortized Gaussian Process Inference and Hidden Markov Models
11:45–12:15	BM	**Yichen Ji** DoSS	Hierarchical Bayesian Copula Model for Probabilistic Population Projection
12:15-13:15	**Lunch (Provided)**		
	Block 3: Mathematical Finance and Actuarial Science		
13:15-13:45	PH	**Brandon Tam** DoSS	Dynamic Pareto Optima in Multi-Period Pure-Exchange Economies
13:45-14:15	PH	**Chenxin Lyu** Waterloo	Adaptive Window Selection for Financial Risk Forecasting
14:15-15:30	**Poster Session**		
	Block 4: Machine Learning		
15:30–16:00	KCT	**Ryan DeWolfe** TMU Mathematics	A Pragmatic Method for Comparing Clusterings with Overlaps and Outliers
16:00–16:30	KCT	**Ziyi Liu** DoSS	Sequential Probability Assignment against Smoothed Adversaries with Unknown Base Measure
16:30	**Closing Remarks**		

Chairs: PC – Pat Chimtanoo; BM - Bernard Miškić; PH – PH Chan; KCT - KC Tsiolis
Note: DoSS – Department of Statistical Sciences

List of Abstracts: Oral Presentations

Block 1: Data Science Theory

Convergence and Optimality of the EM Algorithm Under Multi-Component Gaussian Mixture Models

Bingqing Liu

Department of Statistical Sciences, University of Toronto

Gaussian mixture models (GMMs) are fundamental statistical tools for modeling heterogeneous data. Due to the nonconcavity of the likelihood function, the Expectation-Maximization (EM) algorithm is widely used for parameter estimation of each Gaussian component. Existing analyses of the EM algorithm's convergence to the true parameter focus on either the two-component case or multi-component settings with known mixing probabilities and isotropic covariance matrices.

In this work, we study the convergence of the EM algorithm for multi-component GMMs in full generality. The population-level EM is shown to converge to the true parameter when the smallest separation among all pairs of Gaussian components exceeds a logarithmic factor of the largest separation and the reciprocal of the minimal mixing probabilities. At the sample level, the EM algorithm is shown to be minimax rate-optimal, up to a logarithmic factor. We develop two distinct novel analytical approaches, each tailored to a different regime of separation, reflecting two complementary perspectives on the use of EM. As a byproduct of our analysis, we show that the EM algorithm, when used for community detection, also achieves the minimax optimal rate of misclustering error under milder separation conditions than spectral clustering and Lloyd's algorithm, an interesting result in its own right. Our analysis allows the number of components, the minimal mixing probabilities, the separation between Gaussian components and the dimension to grow with the sample size. Simulation studies corroborate our theoretical findings.

Luis Sierra Muntané

Department of Statistical Sciences, University of Toronto

Low-rank matrix approximation serves as a cornerstone for recovering structured data in fields ranging from medical imaging to recommendation systems. However, its performance hinges on the careful calibration of regularization parameters, a task that remains nontrivial for estimators defined as solutions to regularized problems.

This work explores the use of Efron's degrees of freedom as a principled framework for deriving Stein's unbiased risk estimate (SURE) for spectral shrinkage estimators, while also providing a practical tool for model comparison. By characterizing the differentiability of these estimators, we obtain an intuitive interpretation of the risk formula that links degrees of freedom to an effective notion of matrix rank.

We further extend this framework to a broad class of spectral operators, with particular attention to singular value sparsity and slow decay patterns. Ongoing work also develops an approximation to Tibshirani's search degrees of freedom for hard-thresholding and for sparsity constraints on individual matrix entries, offering a path toward more robust low-rank methods for high-stakes applications.

Block 2: Applied Statistics

Scalable Bayesian Additive Models for Stellar Flare Detection via Amortized Gaussian Process Inference and Hidden Markov Models

Rodrigo Herrera

Department of Statistical Sciences, University of Toronto

Gaussian Processes represent the gold standard for Bayesian time-series modeling, yet their cubic computational complexity remains a barrier for high-cadence datasets. This bottleneck is particularly restrictive when these processes are integrated as latent components within complex likelihood functions, such as Hidden Markov Models, where iterative inference is computationally prohibitive. We introduce a Generative Surrogate framework to overcome these limitations. By utilizing a Variational Autoencoder to learn a compressed representation of a Gaussian Process prior, we map high-dimensional stochastic dependencies into a low-dimensional manifold. This transition effectively reduces inference complexity from cubic to linear. Our approach allows for the seamless integration of non-parametric priors into hybrid architectures without added cost complexity. We demonstrate this on astrophysical light curves, enabling rigorous, large-scale characterization across massive data archives.

Hierarchical Bayesian Copula Model for Probabilistic Population Projection

Yichen Ji

Department of Statistical Sciences, University of Toronto

Population forecasts inform critical decisions in public policy and economic planning, yet existing state-of-the-art methods often underestimate predictive uncertainty by modeling key demographic variables, such as fertility rates and life expectancy, as independent stochastic processes. This work proposes a hierarchical Bayesian framework for probabilistic population forecasting that models the joint dynamics of fertility and mortality across countries, regions, and time. Dependence structure will be represented using copulas, which allow flexible joint modeling while preserving interpretable marginal structures. The hierarchical design enables partial pooling across countries and regions, allowing countries with sparse data to be partially informed by broader regional patterns of demographic change while retaining country-specific trajectories. By producing more faithful joint predictive uncertainty quantification, this work delivers uncertainty-aware population projections that better support evidence-based policy and equitable decision-making for all countries worldwide. This project is still in progress.

Block 3: Mathematical Finance and Actuarial Science

Dynamic Pareto Optima in Multi-Period Pure-Exchange Economies

Brandon Tam

Department of Statistical Sciences, University of Toronto

We study a problem of optimal allocation in a discrete-time multi-period pure-exchange economy, where agents have preferences over stochastic endowment processes that are represented by strongly time-consistent dynamic risk measures. We introduce the notion of dynamic Pareto-optimal allocation processes and show that such processes can be constructed recursively starting with the allocation at the terminal time. We further derive a comonotone improvement theorem for allocation processes, and we provide a recursive approach to constructing comonotone dynamic Pareto optima when the agents' preferences are coherent and satisfy a property that we call equidistribution-preserving. In the special case where each agent's dynamic risk measure is of the distortion type, we provide a closed-form characterization of comonotone dynamic Pareto optima. We illustrate our results in a two-period setting.

Adaptive Window Selection for Financial Risk Forecasting

Chenxin Lyu

Department of Statistics and Actuarial Science, University of Waterloo

Risk forecasts in financial regulation and internal management are calculated through historical data. The unknown structural changes of financial data poses a substantial challenge in selecting an appropriate look-back window for risk modeling and forecasting. We develop a data-driven online learning method, called the bootstrap-based adaptive win-dow selection (BAWS), that adaptively determines the window size in a sequential manner. A central component of BAWS is to compare the realized scores against a data-dependent threshold, which is evaluate based on an idea of bootstrap. The proposed method is applicable to the forecast of risk measures that are elicitable individually or jointly, such as the Value-at-Risk (VaR) and the pair of the VaR and the corresponding Expected Shortfall. Through simulation studies and empirical analyses, we demonstrate that BAWS generally outperforms the standard rolling window approach and the recently developed method of stability-based adaptive window selection, especially when there are structural changes in the data-generating process.

Block 4: Machine Learning

A Pragmatic Method for Comparing Clusterings with Overlaps and Outliers

Ryan DeWolfe

Department of Mathematics, Toronto Metropolitan University

Clustering algorithms are an essential part of the unsupervised data science ecosystem, and extrinsic evaluation of clustering algorithms requires a method for comparing the detected clustering to a ground truth clustering. In a general setting, the detected and ground truth clusterings may have outliers (objects belonging to no cluster), overlapping clusters (objects may belong to more than one cluster), or both, but methods for comparing these clusterings are currently undeveloped. In this note, we define a pragmatic similarity measure for comparing clusterings with overlaps and outliers, show that it has several desirable properties, and experimentally confirm that it is not subject to several common biases afflicting other clustering comparison measures.

Sequential Probability Assignment against Smoothed Adversaries with Unknown Base Measure

Ziyi Liu

Department of Statistical Sciences, University of Toronto

Smoothed online learning has recently been studied as a way to bypass hardness results for the fully adversarial setting, which can be overly pessimistic. In this framework, the adversary is constrained in the sense that contexts are generated by distributions whose densities have to be bounded with respect to some base measure μ. Most prior work makes the strong assumption that μ is known to the learner, with notable exceptions including Block et al. (2024) and Blanchard (2025). In this paper, we study sequential probability assignment (a.k.a. online learning with log loss) with smooth, well-specified data in the more general setting where μ is *unknown*, going beyond the Lipschitz loss studied in Block et al. (2024) and Blanchard (2025). Our main result is a regret upper bound in terms of the *empirical Hellinger entropy*, a notion that has been shown to characterize the complexity of learning with i.i.d. data (Bilodeau et al., 2023). We also prove a matching lower bound showing that our upper bound is essentially tight for a broad range of classes, which also implies a separation in the difficulty of smoothed online learning between regimes where μ is known and where it is unknown.

List of Abstracts: Poster Presentations

Reducing Dimensionality and Multicollinearity in K-mer Data Using Correlation-Based Clustering and Penalized Regression

Alexia Furtado

Department of Mathematics and Statistics, University of Guelph

K-mers are nucleotide sequences derived from DNA and serve as biomarkers for detecting pathogens and antimicrobial resistance. However, k-mer data is high dimensional, sparse and highly collinear, limiting the performance of traditional penalized regression models. To address these challenges, a two-stage framework is proposed. In stage one, pairwise kendall correlations among k-mers are computed and transformed into a pseudo-Euclidean distance to cluster co-occurring k-mers, reducing multicollinearity and dimensionality. In stage two, representative k-mers from each cluster are selected for a penalized regression model where balanced resampling and stability selection are used to address severe class imbalance and enhance robustness. This framework is applied to 156 swine microbiome samples containing over 26,000 k-mers to predict swine type (piglet or sow) and farm type (pasture or conventional), improving model stability, interpretability and preserving biological relevance.

Predicting and improving test-time scaling laws via reward tail-guided search

Muheng Li

Department of Statistical Sciences, University of Toronto

Test-time scaling has emerged as a critical avenue for enhancing the reasoning capabilities of Large Language Models (LLMs). Though the straight-forward "best-of-N" (BoN) strategy has already demonstrated significant improvements in performance, it lacks principled guidance on the choice of N, budget allocation, and multi-stage decision-making, thereby leaving substantial room for optimization. While many works have explored such optimization, rigorous theoretical guarantees remain limited. In this work, we propose new methodologies to predict and improve scaling properties via tail-guided search. By estimating the tail distribution of rewards, our method predicts the scaling law of LLMs without the need for exhaustive evaluations. Leveraging this prediction tool, we introduce Scaling-Law Guided (SLG) Search, a new test-time algorithm that dynamically allocates compute to identify and exploit intermediate states with the highest predicted potential. We theoretically prove that SLG achieves vanishing regret compared to perfect-information oracles, and achieves expected rewards that would otherwise require a polynomially larger compute budget required when using BoN. Empirically, we validate our framework across different LLMs and reward models, confirming that tail-guided allocation consistently achieves higher reward yields than Best-of-N under identical compute budgets. Our code is available at https://github.com/PotatoJnny/Scaling-Law-Guided-search.

Fast computation and marginal density estimation in nonparametric exponential family mixtures

Yan Zhang

Department of Statistical Sciences, University of Toronto

We study the computational and statistical properties of the approximate nonparametric maximum likelihood estimator (NPMLE) for a broad class of exponential family mixture models. This framework includes Gaussian location mixtures and scaled chi-square mixtures as important special cases. We first develop a data compression strategy that reduces the computational cost of the approximate NPMLE to logarithmic order in the sample size. We then show that, for a broad class of approximate NPMLEs, the resulting marginal density estimator attains an almost parametric convergence rate.

From Information to Generative Exponent: Learning Rate Induces Phase Transitions in SGD

KC Tsiolis

Department of Statistical Sciences, University of Toronto

To understand feature learning dynamics in neural networks, recent theoretical works have focused on gradient-based learning of Gaussian single-index models, where the label is a nonlinear function of a latent one-dimensional projection of the input. While the sample complexity of online SGD is determined by the information exponent of the link function, recent works improved this by performing multiple gradient steps on the same sample with different learning rates — yielding a non-correlational update rule — and instead are limited by the (potentially much smaller) generative exponent. However, this picture is only valid when these learning rates are sufficiently large. In this paper, we characterize the relationship between learning rate(s) and sample complexity for a broad class of gradient-based algorithms that encapsulates both correlational and non-correlational updates. We demonstrate that, in certain cases, there is a phase transition from an "information exponent regime" with small learning rate to a "generative exponent regime" with large learning rate. Our framework covers prior analyses of one-pass SGD and SGD with batch reuse, while also introducing a new layer-wise training algorithm that leverages a two-timescales approach (via different learning rates for each layer) to go beyond correlational queries without reusing samples or modifying the loss from squared error. Our theoretical study demonstrates that the choice of learning rate is as important as the design of the algorithm in achieving statistical and computational efficiency.

Discrimination-Insensitive Pricing

Kathleen Miao

Department of Statistical Sciences, University of Toronto

Rendering fair prices for financial, credit, and insurance products is of ethical and regulatory interest. In many jurisdictions, discriminatory covariates, such as gender and ethnicity, are prohibited from use in pricing such instruments. In this work, we propose a discrimination-insensitive pricing framework, where we require the pricing principle to be insensitive to the (exogenously determined) protected covariates, that is the sensitivity of the pricing principle to the protected covariate is zero. We formulate and solve the optimisation problem that finds the nearest (in Kullback-Leibler (KL) divergence) "pricing" measure to the real world probability, such that under this pricing measure the principle is discrimination-insensitive. We call the solution the discrimination-insensitive measure and provide conditions for its existence and uniqueness. In situations when there are more than one protected covariates, the discrimination-insensitive pricing measure might not exist, and we propose a two-step procedure. First, for each protected covariate separately, we find the measure under which the pricing principle becomes insensitivity to that covariate. Second we reconcile these measures through a constrained barycentre model. We provide a close-form solution to this problem and give conditions for existence and uniqueness of the constrained barycentre pricing measure. As an intermediary result, we prove the representation, existence, and uniqueness of the KL barycentre of general probability measures, which may be of independent interest. Finally, in a numerical illustration, we compare our discrimination-insensitive premia and the constrained barycentre pricing measure with recently proposed fair premia from the actuarial literature.