Minghui Yu Memorial Conference

May 1, 2022

Doctoral Student Body
Department of Statistics
Columbia University

About

The 2022 Minghui Yu Memorial Conference, organized by doctoral students in the Statistics Department of Columbia University, will take place on Sunday, May 1st at the Faculty House. Minghui Yu was a doctoral student at the statistics department, who passed away in a tragic accident in the spring of 2008. Since then, doctoral students in the Statistics Department have been organizing a conference each year to honor his memory. The conference will feature talks by doctoral students at the Statistics Department, ranging from those just beginning a research program to those who are about to defend dissertations. In addition to being an occasion to remember our friend and colleague, this event will be an opportunity to learn about exciting new research areas emerging from our department. We would like to thank the Department of Statistics for their continued support.

Location

The conference will be held at the Faculty House (64 Morningside Drive) on the 2nd floor in the Seminar Ballroom. Lunch and dinner will be held in the Presidential Ballroom.

Contact

If you have any questions, please do not hesitate to contact Collin Cademartori at cac2301@columbia.edu.

Keynote - Professor Bin Yu

Bin Yu is Chancellor's Distinguished Professor and Class of 1936 Second Chair in the departments of statistics and EECS at UC Berkeley. She leads the Yu Group which consists of students and postdocs from Statistics and EECS. She was formally trained as a statistician, but her research extends beyond the realm of statistics. Together with her group, her work has leveraged new computational developments to solve important scientific problems by combining novel statistical machine learning approaches with the domain expertise of her many collaborators in neuroscience, genomics and precision medicine. She and her team develop relevant theory to understand random forests and deep learning for insight into and guidance for practice.

She is a member of the U.S. National Academy of Sciences and of the American Academy of Arts and Sciences. She is Past President of the Institute of Mathematical Statistics (IMS), Guggenheim Fellow, Tukey Memorial Lecturer of the Bernoulli Society, Rietz Lecturer of IMS, and a COPSS E. L. Scott prize winner. She holds an Honorary Doctorate from The University of Lausanne (UNIL), Faculty of Business and Economics, in Switzerland. She has recently served on the inaugural scientific advisory committee of the UK Turing Institute for Data Science and AI, and is serving on the editorial board of Proceedings of National Academy of Sciences (PNAS).

Schedule

8:00AM - 9:15AM	Breakfast
9:15AM - 9:20AM	Opening remarks by Tian Zheng
	Session #1
9:20AM - 9:40AM	*Spatio-temporal Analysis of Embolism Formation in Leaves using Spatial Survival Analysis*, Diane Lu
9:40AM - 10:00AM	*Multi-animal Pose Estimation*, Ari Blau
10:00AM - 10:20AM	*Estimating the Incidence of Sexual Assault on College Campuses*, Casey Bradshaw
10:20AM - 10:50AM	Coffee and tea break
	Session #2
10:50AM - 11:10AM	*Dependent Stopping Times*, Alejandra Quintos
11:10AM - 11:30AM	*Longitudinal Latent Space Model for Network Data*, Jiajin Sun
11:30AM - 11:50AM	*Limits and Optimal Portfolios of Semistatic Trading Strategies*, Long Zhao
11:50AM - 1:00PM	Lunch
1:00PM - 2:00PM	**Keynote:** *Predictability, Stability and Causality with a Case Study to Find Genetic Drivers of a Heart Disease*, Bin Yu
	Session #3
2:00PM - 2:20PM	*Transfer Learning under High-dimensional Generalized Linear Models*, Ye Tian
2:20PM - 2:40PM	*Data Augmentation for Compositional Data*, Elliott Gordon Rodriguez
2:40PM - 3:00PM	*Variable Selection Using Kernel Partial Correlation Coefficient*, Zhen Huang
3:00PM - 3:30PM	Coffee and tea break
	Session #4
3:30PM - 3:50PM	*Energy Statistics for Clustering in Time Series*, Leon Fernandes
3:50PM - 4:10PM	*Fluctuations in Random Field Ising Models on a Regular Graph*, Seunghyun Lee
4:10PM - 4:30PM	*Posterior Entropy and Bayesian Model Expansion*, Collin Cademartori
4:30PM - 4:50PM	*Nuances in Margin Conditions Determine Gains in Active Learning*, Gan Yuan
6:00PM-8:00PM	Dinner

Student Talks

Diane Lu

Spatio-temporal Analysis of Embolism Formation in Leaves using Spatial Survival Analysis

When there are too many xylem vein vessels become air-filled (i.e. "embolized"), the plant might experience irreversible hydraulic failure. The topic of embolism formation has typically been studied by using the vulnerability curve, which only captures temporal information. With the data collected using the optical method, we're able to perform spatio-temporal analysis. In our study, we try to understand whether there's spatial dependence between embolism formation through spatial survival analysis.

Ari Blau

Multi-animal Pose Estimation

Multi-animal pose estimation is essential for studying animals' social behaviors in neuroscience and neuroethology. Advanced approaches have been proposed to support multi-animal estimation. However, these models rarely exploit unlabeled data during training, even though real world applications have exponentially more unlabeled frames than labeled frames. Manually adding dense annotations for a large number of images or videos is costly and labor-intensive, especially for multiple instances. Given these deficiencies, we propose a novel semi-supervised architecture for multi-animal pose estimation, leveraging the abundant structures pervasive in unlabeled frames in behavior videos to enhance training, particularly in the context of sparsely-labeled problems. The resulting algorithm provides superior multi-animal pose estimation results on three animal experiments compared to existing baselines, and exhibits more predictive power in sparsely-labeled data regimes.

Casey Bradshaw

Estimating the Incidence of Sexual Assault on College Campuses

Each year, US colleges and universities are required to disclose the number of reported sexual assaults on their campuses. However, sexual assault is widely believed to be underreported, and the number of reported assaults could arise from any combination of reporting rate and true number of assaults. In this talk, I will discuss Bayesian modeling strategies for disentangling those two variables. With certain structural assumptions and judicious use of prior information, we can identify plausible values for the total number of assaults occurring in a given year, as well as plausible reporting rates. Such estimates improve the interpretability of campus crime statistics and may help inform policy decisions regarding campus initiatives for sexual assault prevention and awareness.

Alejandra Quintos

Dependent Stopping Times

Stopping times are used in applications to model random arrivals. A standard assumption in many models is that the stopping times are conditionally independent, given an underlying filtration. This is a widely useful assumption, but there are circumstances where it seems to be unnecessarily strong. We use a modified Cox construction, along with the bivariate exponential introduced by Marshall & Olkin (1967), to create a family of stopping times, which are not necessarily conditionally independent, allowing for a positive probability for them to be equal. We also present a series of results exploring the special properties of this construction.

Jiajin Sun

Longitudinal Latent Space Model for Network Data

In the big data era a lot of network data comes with additional time stamp information of the event history on the observed edges. As many successful methods have been established for analyzing static networks, there is a growing interest now in modeling longitudinal network data and especially, how to incorporate the event time information. In this work we propose a longitudinal latent space model, which consists of time-varying individual-specific baselines and parameter-of-interest latent states that doesn't change with time. We show theoretically that the error rate of the penalized maximum profile likelihood estimator of the the latent positions is of order $O_P(1/T)$, which indicates we could accumulate the information shared across longitudinal observations. We demonstrate the good performance of our method in the New York Citi Bike dataset. Using the transport history among bike stations we manage to retrieve the geographical positions of the stations, and the estimated intensity baselines show peaks in rush hours as expected.

Long Zhao

Limits and Optimal Portfolios of Semistatic Trading Strategies

Our talk consists of two parts: First, we show that pointwise limits of semistatic trading strategies in discrete time are again semistatic strategies. The analysis is carried out in full generality for a two-period model, and under a probabilistic condition for multi-period, multi-stock models. Our result contrasts with a counterexample of Acciaio, Larsson and Schachermayer, and shows that their observation is due to a failure of integrability rather than instability of the semistatic form. Second, in a two-period financial market with semistatic trading, we study the so-called martingale Schrodinger bridge Q; that is, the minimal-entropy martingale measure among all models calibrated to option prices. This minimization is shown to be in duality with an exponential utility maximization over semistatic portfolios. Under a technical condition on the physical measure P, we show that an optimal portfolio exists and provides an explicit solution for Q.

Ye Tian

Transfer Learning under High-dimensional Generalized Linear Models

In this work, we study the transfer learning problem under high-dimensional generalized linear models (GLMs), which aim to improve the fit on *target* data by borrowing information from useful *source* data. Given which sources to transfer, we propose a transfer learning algorithm on GLM, and derive its ℓ_1/ℓ_2-estimation error bounds as well as a bound for a prediction error measure. The theoretical analysis shows that when the target and source are sufficiently close to each other, these bounds could be improved over those of the classical penalized estimator using only target data under mild conditions. When we don't know which sources to transfer, an *algorithm-free* transferable source detection approach is introduced to detect informative sources. The detection consistency is proved under the high-dimensional GLM transfer learning setting. We also propose an algorithm to construct confidence intervals of each coefficient component, and the corresponding theories are provided. Extensive simulations and a real-data experiment verify the effectiveness of our algorithms. We implement the proposed GLM transfer learning algorithms in a new R package `glmtrans`, which is available on CRAN.

Elliott Gordon Rodriguez

Data Augmentation for Compositional Data

Data augmentation plays a key role in modern machine learning pipelines. However, knowing what constitutes a useful data augmentation is highly domain-dependent. While numerous augmentation strategies have been studied in the context of computer vision and natural language processing, less is known for other data modalities. The goal of our work is to extend the success of data augmentation to compositional data. Compositional data refers to simplex-valued data, which is of particular interest in the context of microbiology and high-throughput genetic sequencing. Motivated by the special characteristics of these data, we propose bespoke augmentation strategies, which provide significant performance improvements across a range of microbiome studies.

Zhen Huang

Variable Selection Using Kernel Partial Correlation Coefficient

Suppose we have a response Y and 1,000 covariates Xi, and Y is a function of the first three covariates plus noise. Y and Xi may not even be Euclidean. How can we identify the "true" three predictive variables exactly given 200 iid observations? A variable selection algorithm, which we call KFOCI, can do the job even without the need to specify the number of variables to select, achieving much superior performance compared with its predecessors, and is provably consistent under sparsity assumption. KFOCI is an application of the kernel partial correlation coefficient (KPC) we propose, which is a number between 0 and 1 measuring the strength of conditional dependence–the KPC between two random variables Y and Z given a third variable X is 0 if and only if Y is conditionally independent of Z given X, and 1 if and only if Y is a measurable function of Z and X. Given the predictors that have already been selected, KFOCI selects the next predictor Xi such that the sample KPC between Y and Xi given the selected predictors is maximized, and stops when all such sample KPC are negative. Both KPC and KFOCI are easily accessible through our package KPC available on CRAN.

Leon Fernandes

Energy Statistics for Clustering in Time Series

We consider the problem of clustering multivariate time series based on the lagged stationary distributions of its components. The dissimilarity or distance between each pair of components is measured using energy distance, developed by Székely and Rizzo, 2013 (J. Statist. Plann. Inference 143 1249-1272); we obtain the consistency and asymptotic distributions of the energy statistics in a general time series setting. These results are based on joint work with Richard Davis and Konstantinos Fokianos.

Seunghyun Lee

Fluctuations in Random Field Ising Models on a Regular Graph

Random field Ising models originate from statistical physics but are also useful in statistics, especially to model binary dependent data with covariates. In this talk, we explore two fundamental properties of this model: the law of large numbers and the central limit theorem. First, we prove the law of large numbers for the average magnetization. This is similar to the result from the usual

Ising model. Next, we extend this to fluctuations and analyze both quenched and annealed limiting behaviors. Our results show that we need an additional centering for the CLT, and the random fields lead to a higher critical temperature. Also, we anticipate that the order of fluctuations at criticality also changes. This is joint work with Sumit Mukherjee and Nabarun Deb.

Collin Cademartori

Posterior Entropy and Bayesian Model Expansion

What happens as we add predictors to a simple linear regression model with fixed data size? Once the number of predictors outstrips the number of data points, the model becomes non-identified. In a Bayesian setting, this creates a posterior distribution which exactly recovers the prior in the non-identified directions. We consider a generalization of this phenomenon by using a comparison between posterior and prior entropy to quantify the degree of identification, and we seek constraints on the behavior of this quantity under more general conditions of model expansion. Our main results (i) give conditions under which sufficient declines in the curvature of the likelihood translate to guaranteed increases in expected posterior entropy and (ii) provide a bound on the relevant measure of curvature within a general framework for model expansion.

Gan Yuan

Nuances in Margin Conditions Determine Gains in Active Learning

We consider nonparametric classification with smooth regression functions, where it is well known that notions of margin in E[Y |X] determine fast or slow rates in both active and passive learning. Here we elucidate a striking distinction between the two settings. Namely, we show that some seemingly benign nuances in notions of margin—somehow involving the uniqueness of the Bayes classifier, and which have no apparent effect on rates in passive learning—determine whether or not any active learner can outperform passive learning rates. In particular, for Audibert Tsybakov's margin condition (allowing general situations with non-unique Bayes classifiers), no active learner can gain over passive learning in commonly studied settings where the marginal on X is near uniform. Our results thus negate the usual intuition from past literature that active rates should generally improve over passive rates in nonparametric settings.

About Minghui Yu

Minghui was born in Shandong, China in 1983. In 2002, he entered the Special Class for the Gifted Young at the University of Science and Technology of China (USTC), one of the most prestigious universities in China. Minghui possessed the rare quality of being not only smart, but also diligent, versatile, modest and easy-going. He was the type of friend who would stand by you no matter the situation. Minghui breezed through the challenging undergraduate program at USTC, ranking at the top of his class. Minghui was well liked by his fellows students having served as the class president from his sophomore year. Although under enormous academic pressure, he still found time to organize a series of student activities, such as hiking, art performances, and athletic contests for his fellow students.

After graduating summa cum laude in 2006 from USTC, Minghui entered the PhD program at the Physics Department of Columbia University. After one year, he transferred to the doctorate program in statistics. During his time at Columbia, Minghui served as the public relations head of the Columbia University's Chinese Students and Scholars Association (2007-2008), and was a member of the Columbia Chinese Basketball Association and the Columbia Graduate Student Consulting Club. His biography on the CUCSSA website mentioned his love of "movies, photography and delicacies". Minghui described himself in his blog as a boy who wants to combine art and science together.

On April 4, 2008, after attending a student- organized conference, Minghui escorted his girlfriend home on the west side of campus. On his return, he was accosted as he was crossing 122nd and Broadway and in his attempt to flee, he was struck by an automobile on Broadway. Minghui was taken to St. Luke's Hospital where he passed away a short time later.