2019 Minghui Yu Memorial Conference

April 27th 2019

About

The 2019 Minghui Yu Memorial Conference, organized by doctoral students at the Statistics Department of Columbia University, will take place on Saturday, April 27th 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 at 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 1st Floor in Garden Room 2. Lunch and dinner will be held in the Presidential Ballroom.

Contact

If you have any questions, please do not hesitate to contact Andrew Davison at ad3395@columbia.edu.

Timetable

8:55am-9:00am	Opening remarks by Richard Davis
Session 1 (Chair: Cindy Rush)	
9:00am-9:20am	*Spectral Embedding of Optimal Weighted Adjacency Matrix (SE-OWAM)*, Sihan Huang
9:20am-9:40am	*Overlapping Community Detection in Networks*, Alessandro Grande
9:40am-10:00am	*Asymptotic analysis for embeddings of exchangeable graphs*, Andrew Davison
10:00am-10:20am	*Large-scale Spatial Predictive Modeling with Applications to Ecological Remote Sensing Data*, Chengliang Tang
10:20am-10:40am	Coffee break
Session 2 (Chair: Tian Zheng)	
10:40am-11:00am	*A penalized matrix decomposition for denoising and compression of functional imaging data*, Ian Kinsella
11:00am-11:20am	*A Spike+Gamma model for post-deconvolved calcium imaging traces*, Ding Zhou
11:20am-11:40am	*Towards Theoretically-Founded Learning-Based Denoising*, Wenda Zhou
Keynote I	
11:50am-12:30pm	*Diversification, Volatility, and Surprising Alpha*, Johannes Ruf
12:30pm-1:30pm	Lunch
Keynote II	
1:30pm-2:10pm	*Estimating spillovers using imprecisely measured networks*, Tyler McCormick
Session 3 (Chair: Sumit Mukherjee)	
2:20pm-2:40pm	*Fluctuations in the mean-field Ising model*, Nabarun Deb
2:40pm-3:00pm	*Optimal confidence bands under shape restriction in multi-dimensions*, Pratyay Datta
3:00pm-3:20pm	*Multivariate quantiles and ranks using optimal transportation: Consistency and pointwise rate of convergence*, Promit Ghosal
3:20pm-3:40pm	Coffee break

Session 4 (Chair: Bodhi Sen)	
3:40pm-4:00pm	*From Cox Processes to Fair Micro finance Loan Rates*, Alejandra Quintos Lima
4:00pm-4:20pm	*Using High Dimensional Chemical Abundance Data to Better Understand Milky Way Disk Formation and Evolution*, Bridget Ratcliffe
4:20pm-4:40pm	*Latent feature extraction for process data*, Zhi Wang
4:40pm-5:00pm	*Stacking for multimodal posterior distributions*, Yuling Yao
6:00pm	Conference dinner

Keynote Presentations

We are excited to have Johannes Ruf and Tyler McCormick as our keynote speakers this year. Johannes is currently an Associate Professor in the Department of Mathematics at the London School of Economics, with his research focusing on stochastic analysis and its applications to mathematical finance. Tyler is currently an Associate Professor of Statistics and Sociology at the University of Washington, with some of his research interests being in the analysis of social networks and the estimation of demographic indicators with sparsely sampled data. Both were part of the organizing committee for the inaugural Minghui conference.

Diversification, Volatility, and Surprising Alpha, Johannes Ruf

It has been widely observed that capitalization-weighted indexes can be beaten by surprisingly simple, systematic investment strategies. Indeed, in the U.S. stock market, equal-weighted portfolios, random-weighted portfolios, and other naïve, nonoptimized portfolios tend to outperform a capitalization-weighted index over the long term. This outperformance is generally attributed to beneficial factor exposures. Here, we provide a deeper, more general explanation of this phenomenon by decomposing portfolio log-returns into an average growth and an excess growth component. Using a rank-based empirical study we argue that the excess growth component plays the major role in explaining the outperformance of naïve portfolios. In particular, individual stock growth rates are not as critical as is traditionally assumed.

Estimating spillovers using imprecisely measured networks, Tyler McCormick

In many experimental contexts, whether and how network interactions impact the outcome of interest for both treated and untreated individuals are key concerns. Networks data is often assumed to perfectly represent these possible interactions. This paper considers the problem of estimating treatment effects when measured connections are, instead, a noisy representation of the true spillover pathways. We show that existing methods, using the potential outcomes framework, yield biased estimators in the presence of this mismeasurement. We develop a new method, using a class of mixture models, that can account for missing connections. We check our method's performance by simulating experiments on network data from villages in rural India. Finally, we use data from a previously published study to show that estimates using our method are more robust to the choice of network measure that currently available approaches. This is joint work with Morgan Hardy (NYU- Abu Dhabi), Rachel Heath (University of Washington), and Wesley Lee (University of Washington). (See https://arxiv.org/abs/1904.00136 for details.)

Student Abstracts

Optimal confidence bands under shape restriction in multi-dimensions, Pratyay Datta

In this talk we propose to find confidence bands for regression function in multi-dimension. For simplicity we look at standard continuous white noise regression model in the d-dimension hypercube. We use the multidimensional multiscale statistic (which is an extension of the multiscale statistic proposed by Dümbgen(2001)) to construct confidence bands for many shape restricted classes e.g. multivariate isotonic, multivariate convex etc. Our confidence bands are shown to have guaranteed finite-sample coverage probability. Our proposed confidence bands are also shown to be adaptive and optimal (in an appropriate sense) with respect to the smoothness of the underlying regression function. Moreover, these bands are adaptive to the intrinsic dimension (as opposed to the ambient dimension d) of the regression function. We have also constructed asymptotic confidence bands for the "completely monotone" (functions that have a density with respect to the Lebesgue measure) functions and shown the rate of convergence for our confidence band to be dimension free. We have also done the same under the highly useful additive model constraints.

Asymptotic analysis for embeddings of exchangeable graphs, Andrew Davison

Tasks involving the analysis of network data are frequent within statistics. For node classification and link prediction, state-of-the-art performance is frequently achieved via the application of machine learning methods to a learnt embedding of the network into a lower-dimensional Euclidean space. These embeddings are frequently learnt via random-walk based approaches - popularized by DeepWalk (Perozzi, Al-Rfou and Skiena, 2014) and node2vec (Grover and Leskovec, 2016) - where an empirical risk is formed via averaging over losses created via performing random walks on the network graph. While these methods are very successful empirically, there is currently no theoretical understanding of what the embedding vectors represent, the role of the sampling scheme and how this translates to downstream performance. I'll discuss some theoretical progress heading in these directions under exchangability assumptions of the underlying graph. This is based off joint work with Morgane Austern.

Fluctuations in the mean-field Ising model, Nabarun Deb

In a recent work of Basak and Mukherjee (2017) it was shown that whenever the underlying graph is "approximately" regular, the scaled log-normalizing constant of the Ising model on that graph has a universal limit whenever the average degree diverges. Extending their argument one can show that for such Ising models the sum of spins has a universal weak limit. A natural follow up question is whether the fluctuations for the sum of spins exhibit similar universal behavior.

This talk studies fluctuations of the sum of spins for such Ising models at all temperatures (high, low and critical), and show that whenever the average degree grows faster than $\sqrt{n}$

(n=number of vertices of the graph), the sum of spins have a universal limit distribution. Moreover, we show that the $\sqrt{n}$ condition is tight. Finally, we study the fluctuations of a conditionally centered sum of spins, in which case the universality extends to all regular graphs with degree going $+\infty$, thereby covering the entire asymptotic regime considered in Basak-Mukherjee. Finally, we use our results to derive interesting consequences in parameter estimation in Ising models.

Multivariate quantiles and ranks using optimal transportation: Consistency and pointwise rate of convergence, Promit Ghosal

Quantiles and ranks are important tools in nonparametric statistics. We consider multivariate quantiles and ranks using optimal transportation theory. Under some minor assumption, we show a Glivenko-Cantelli type result for our empirical rank function and an almost sure convergence result on compacts for the quantile function. In a special case, we find the local rate of convergence for the empirical quantiles and ranks functions. If time permits, we will discuss some applications. This is based on a joint work with Prof. Bodhisattva Sen.

Overlapping Community Detection in Networks, Alessandro Grande

A fundamental problem in network science is how to detect those communities in which networked systems seem to divide naturally. Early efforts at community detection focused on disjoint communities but, as argued in many recent works, in social and biological networks communities tend to overlap. I present a method for overlapping community detection based on a Bayesian extension of a popular model by Ball, Karrer, and Newman (2011). In this new setting, we can run approximate posterior inference via a coordinate ascent algorithm, and then scale it up to massive networks via stochastic variational inference. I investigate the performance of this method on a wide range of simulated networks.

Spectral Embedding of Optimal Weighted Adjacency Matrix (SEOWAM), Sihan Huang

Nowadays, tons of papers have discussed about the community detection task for a single network, but just a few papers focus on how to utilize the extra information provided by multilayer network, which is actually a very common problem in the real world. For a Multilayer Stochastic Block Model (MSBM), we studied the distribution of its spectrum and eigenvectors of the linear weighted adjacency matrix, and derived the optimal weights in the sense of clustering accuracy. We linked eigenvalue gap with clustering error and proposed an effective algorithm to obtain the empirical optimal weights. This paper provides a comprehensive understanding of the spectral clustering method on MSBM and minimizes the error rate under the frame of linear combination.

A penalized matrix decomposition for denoising and compression of functional imaging data, Ian Kinsella

Calcium and voltage imaging – tools for recording from large neural populations with single-cell resolution — play a critical role in systems neuroscience. In order to analyze data obtained with these imaging modalities, the observed fluorescence must first be decomposed into contributions from individual neurons. State of the art methods addressing this challenge are based on non-negative matrix factorization (NMF); these approaches are highly effective when good initializations are available, but can break down e.g. in low-SNR settings. The aforementioned challenge coupled with the increasing the size of imaging datasets (with scales of TB/hour in some cases) present a barrier to routine processing and data sharing, slowing progress in reproducible research. We introduce a robust and scalable approach for compressing and denoising functional imaging data based on a spatially-localized penalized matrix decomposition (PMD). We have applied the method to a wide range of functional imaging data (including one-photon, two-photon, three-photon, widefield, somatic, axonal, dendritic, calcium, and voltage imaging datasets), with no adjustment of hyperparameters. In all cases, we observe 2-4x increases in SNR and compression rates of 20-300x with minimal visible loss of signal. Furthermore, we demonstrate that this in turn boosts the efficiency of subsequent applications of NMF-based demixing methods.

From Cox Processes to Fair Micro finance Loan Rates, Alejandra Quintos Lima

In this talk, we will provide a brief introduction to the Cox Process (a general construction of a compensated Poisson Process) as it is one of the main tools used in our reduced form credit risk model for micro finance loans. The lending mechanism that we are working with is that of group lending, which is defined as a loan issued by one lender to all members in a group where all of them are charged the same lending rate and, if any member defaults, no one in the group can get future loans. We will present the two challenges we are currently facing: determining bounds for the fair lending rate and the optimal size of the group.

Using High Dimensional Chemical Abundance Data to Better Understand Milky Way Disk Formation and Evolution, Bridget Ratcliffe

The chemical composition of heavy elements in a star varies depending on the evolution of the gas from which it formed. Therefore a survey of stellar chemical abundances should provide information about the poorly understood formation and evolution of the Milky Way disk. To date, the narrative in the field has centered around the separation of stars in the Milky Way disk into two clusters, the high and low α sequences. However, this clustering is performed primarily by eye, using only 2 dimensional abundance data ([α/Fe] vs [Fe/H]). Now that higher dimensional chemical abundance data is available, we aim to produce a more informative clustering of stars by using agglomerative hierarchical clustering. Through exploration of 19 chemical abundance ratios of 27,135 Red Clump stars taken from the APOGEE survey, we show that some stars previously classified in the high α sequence more closely resemble low α sequence stars in the higher dimensional space. Additionally, we investigate the underlying dimensionality of the data through the use of Principal Component Analysis (PCA) and Sparse PCA. We can relate the way different elements contribute towards the first two principal components with how the elements are produced. Further, we briefly discuss nonlinear dimension reduction techniques.

Large-scale Spatial Predictive Modeling with Applications to Ecological Remote Sensing Data, Chengliang Tang

Climate change is anticipated to have profound implications for the long- term forest ecosystem resilience and increase carbon fluxes to the atmosphere. Our understanding of these critical problems have been limited by conventional plot-level ground-based observations. Recently developed remote sensing techniques such as stereo photos provide large-scale high-resolution data that would allow researchers to carry out ecological surveys of forest at an unprecedented scale. In this work, we propose a data science workflow to predict palm tree density on forested landscapes in Puerto Rico by applying machine learning tools to data from imaging and remote sensing technologies, combined with ground observation data from field plots. Also, we perform elementary data analysis based on the palm density prediction to reveal its link with other environmental factors, and verify existing ecological theories about landscape characteristics.

Latent feature extraction for process data, Zhi Wang

Computer simulations have become a popular tool of assessing complex skills such as problem-solving skills. The log files of computer-based items record the entire process of solving the items for each respondent. The recorded process data is very diverse, noisy, and in a non-standard format. Few general methods have been developed for exploiting the information contained in process data. We propose automatic methods, including multidimensional scaling and sequence-to-sequence autoencoder, to extract the latent features and explore some latent structures through predictive modeling. It does not require prior knowledge of the items and human-computers interaction patterns.

Stacking for multimodal posterior distributions, Yuling Yao

When working with multimodal posterior distributions, MCMC algorithms can have difficulty moving between modes, and default variational or mode-based approximate inferences can understate posterior uncertainty. And, even if the most important modes can be found, it is difficult to evaluate their relative weights in the posterior, which requires computing the integral of the posterior in the neighborhood of each mode. Here we propose an alternative approach, using parallel runs of MCMC, variational, or mode-based inferences to hit as many modes as possible, and then using *Bayesian stacking* to weight the set of simulations at each mode. Bayesian stacking is a method for constructing a weighted average of distributions so as to minimize cross-validated prediction errors. The result from stacking is not necessarily equivalent, even asymptotically, to fully Bayesian inference, but it serves many of the same goals.

A Spike+Gamma model for post-deconvolved calcium imaging traces, Ding Zhou

Calcium imaging is a critical tool for measuring the activity of large neural populations. Much effort has been devoted to developing "pre-processing" tools, addressing the important issues of e.g., motion correction, denoising, compression, demixing, and deconvolution. However, computational modeling of deconvolved calcium signals (i.e., the estimated activity extracted by a pre-processing pipeline) is just as critical for interpreting calcium measurements. Surprisingly, these issues have to date received significantly less attention. To fill this gap, we examine the statistical properties of the deconvolved activity estimates, and propose several density models for these random signals. These models include a Spike+Gamma model, which characterizes the calcium responses as a mixture of a gamma distribution and a point mass ("spike") which serves to model zero responses. We apply the resulting models to neural encoding and decoding problems. We find that the Spike+Gamma model outperforms simpler models (e.g., Poisson or Bernoulli models) in the context of both simulated and real neural data, and can therefore play a useful role in bridging calcium imaging analysis methods with tools for analyzing activity in large neural populations.

Towards Theoretically-Founded Learning-Based Denoising, Wenda Zhou

Denoising a stationary process corrupted by additive white Gaussian noise is a fundamental problem in statistical signal processing. For general distributions, theoretically-founded computationally-efficient denoising algorithms are yet to be found. Starting from a Bayesian setup, where the source distribution is fully known, we propose the quantized MAP (Q-MAP) denoiser, and analyze its asymptotic performance. A key advantage of the Q-MAP denoiser is that it highlights the key properties of the source that are relevant to its denoising. This property dramatically reduces the computational complexity of approximating the solution of the Q-MAP denoiser and inspires a novel learning-based denoiser.

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.