# Minghui Yu Memorial Conference 2018

## April 28, 2018

# About

The 2018 Minghui Yu Memorial Conference, organized by doctoral students at the Statistics Department of Columbia University, will take place on Saturday, April 28th at the Faculty House in the Presidential Ballroom. 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) in the Presidential Ballroom. The dinner will be held at Flat Top (1255 Amsterdam Avenue, New York, NY 10027).

## Contact

If you have any questions please do not hesitate to email florian@stat.columbia.edu

## Schedule

| | |
|---|---|
| 9:00am–9:25am | Breakfast |
| 9:25am–9:30am | Opening Remarks |
| Session 1 | Chair: Peter Orbanz |
| 9:30am–9:45am | Rishabh Dudeja—High Dimensional Optimization Landscape of the Regularized MLE for a Symmetric Mixture of Two Gaussians |
| 9:45am–10:00am | Wenda Zhou—Compressibility and generalization in large-scale deep learning |
| 10:00am–10:15am | Timothy Jones—Aggregated Relational Counts for Citation Networks |
| 10:15am–10:30am | Jing Wu—Modeling Sporadic Event Dynamics with Markov-Modulated Hawkes Processes |
| 10:30am–10:45am | Morgane Austern—TBD |
| 10:45am–11:00am | Coffee Break |
| Session 2 | Chair: Yang Feng |
| 11:00am–11:15am | Milad Bakshizadeh—Compressive Phase Retrieval of Structured Signal |
| 11:15am–11:30am | Sihan Huang—Pairwise Covariates-adjusted Block Model for Community Detection |
| 11:30am–11:45am | David Hirshberg—Augmented Linear Minimax Estimation |
| 11:45am–12:00am | Florian Stebegg—Robust Pricing and Hedging around the Globe |
| 12:00pm–1:00pm | Lunch |
| 1:00pm–2:00pm | Keynote Talk - Ed George |
| 2:00pm–2:15pm | Coffee Break |
| Session 3 | Chair: Jingchen Liu |
| 2:15pm–2:30pm | Gabriel Loaiza Ganem—Dimensionality Reduction for Point Process Data |
| 2:30pm–2:45pm | Guanhua Fang—Latent Class Models for Finding Co-occurrent Patterns in Process Data |
| 2:45pm–3:00pm | Hok Kan Ling—A proportional intensity model with random effects for process data |
| 3:00pm–3:15pm | Zhi Wang—Improving approximation via adaptive weighting |
| 3:15pm–3:30pm | Peter Lee—Challenges in spike sorting and YASS: Yet Another Spike Sorter |
| 3:30pm–3:45pm | Coffee Break |
| Session 4 | Chair: Cynthia Rush |
| 3:45pm–4:00pm | Adji Bousso Dieng—Augment and Reduce: Stochastic Inference for Large Categorical Distributions |
| 4:00pm–4:15pm | Yixin Wang—Frequentist Consistency of Variational Bayes |
| 4:15pm–4:30pm | Yuling Yao—Yes, but Did It Work?: Evaluating Variational Inference |
| 4:30pm–4:45pm | Lydia Hsu—Variable Selection in Graphical Models |
| 4:45pm–5:00pm | Jonathan Auerbach—How realistic is the mile-high skyscraper? |
| 5:45pm | Dinner - Flat Top |

# Keynote - Ed George

We are excited to have Ed George as our keynote speaker this year. Prof. George is currently Universal Furniture professor of Statistics at the Wharton School of the University of Pennsylvania. He started his education as an undergraduate at Cornell University and later earned his PhD in Statistics at Stanford University before becoming a Professor at the University of Chicago. He also held positions at UT Austin over the course of his career. His broad contribution to statistical research has been recognized by numerous professional organizations, namely he is an elected fellow of IMS, ASA and ISBA.

## Talk - Bayesian Penalty Mixing with the The Spike and Slab Lasso

Despite the wide adoption of spike-and-slab methodology for Bayesian variable selection, its potential for penalized likelihood estimation has largely been overlooked. We bridge this gap by cross-fertilizing these two paradigms with the Spike-and-Slab Lasso, a procedure for simultaneous variable selection and parameter estimation in linear regression. A mixture of two Laplace distributions, the Spike-and-Slab Lasso prior induces a new class of self-adaptive penalty functions that arise from a fully Bayes spike-and-slab formulation, ultimately moving beyond the separable penalty framework. A virtue of these non-separable penalties is their ability to borrow strength across coordinates, adapt to ensemble sparsity information and exert multiplicity adjustment. With a path-following scheme for dynamic posterior exploration, efficient EM and coordinatewise implementations, the fully Bayes penalty is seen to mimic oracle performance, providing a viable alternative to cross-validation. Further elaborations of the Spike-and-Slab Lasso for fast Bayesian factor analysis illuminate its broad potential. (This is joint work with Veronika Rockova).

# Student Abstracts

## Jonathan Auerbach—How realistic is the mile-high skyscraper?

(joint work with Phyllis Wan)

Half of the world's population lives in cities, up from a third fifty years ago, and this proportion is projected to rise to two-thirds by 2050. City planners have proposed super-tall skyscrapers to accommodate the increase,some reaching a mile into the sky. that's nearly twice the current tallest skyscraper! We investigate whether the mile-high skyscraper is a realistic goal for city planners, and how many residents planners should expect such a building to hold.

## Morgane Austern—TBD

TBD

## Milad Bakshizadeh—Compressive Phase Retrieval of Structured Signal

(joint work with Arian Maleki, Shirin Jalali)

Compressive phase retrieval is the problem of recovering a structured vector $x \in \mathbb{C}^n$ from its phaseless linear measurements. A compression algorithm aims to represent structured signals with as few bits as possible. As a result of extensive research devoted to compression algorithms, in many signal classes, compression algorithms are capable of employing sophisticated structures in signals and compress them efficiently. This raises the following important question: Can a compression algorithm be used for the compressive phase retrieval problem? To address this question, COmpressive PhasE Retrieval (COPER) optimization is proposed, which is a compression-based phase retrieval method. For a family of compression codes with rate-distortion function denoted by $r(\delta)$, in the noiseless setting, COPER is shown to require slightly more than $\lim_{\delta \to 0} r(\delta) \log(1/\delta)$ observations for an almost accurate recovery of x.

## Adji Bousso Dieng—Augment and Reduce: Stochastic Inference for Large Categorical Distributions

Categorical distributions are ubiquitous in machine learning, e.g., in classification, language models, and recommendation systems. They are also at the core of discrete choice models. However, when the number of possible outcomes is very large, using categorical distributions becomes computationally expensive, as the complexity scales linearly with the number of outcomes. To address this problem, we propose augment and reduce (A&R), a method to alleviate the computational complexity. A&R uses two ideas: latent variable augmentation and stochastic variational inference. It maximizes a lower bound on the marginal likelihood of the data. Unlike existing methods which are specific to softmax, A&R is more general and is amenable to other categorical models, such as multinomial probit. On several large-scale classification problems, A&R provides a tighter bound on the marginal likelihood and has better predictive performance than existing approaches.

## Rishabh Dudeja—High Dimensional Optimization Landscape of the Regularized MLE for a Symmetric Mixture of Two Gaussians

We consider the ridge regularized maximum likelihood optimization problem for a symmetric mixture of two gaussians in $p$ dimensions using $n$ data points. Recent results have shown that even though this problem is non-convex, once $n = O(p \log(p))$ data points are sampled, the optimization problem is tractable with a unique local maximum (up to sign symmetry) which is within the rate-optimal distance from the true parameter. We complement these results by studying this problem in the high dimensional asymptotic when $n, p \to \infty$ such that the aspect ratio $n/p \to \delta$. Using non-rigorous techniques from statistical physics we characterize the phase transitions in the landscape of this problem. Depending on the sampling ratio, cluster separation and the regularization parameter the landscape can either be a) convex or b) non-convex with unique local maximum or c) non-convex with numerous stationary points. We further show that by setting the regularization appropriately it is possible to completely avoid the third phase. Finally we also characterize the performance of an Approximate Message Passing algorithm in each of these cases.

## Guanhua Fang—Latent Class Models for Finding Co-occurrent Patterns in Process Data

Process data, a temporal ordered data with irregular structure, is of great interest for research analysis due to its massive underlying information. Each process is a collection of multi-type events along with time stamps, recording how an

individual performs or reacts in certain time period. For example, imagine all the possible events that could occur when a student takes online test, including clicking, searching, scrolling and filtering multiple candidate options. As opposed to traditional cross-sectional response data,process data entails much more features and patterns which could be useful for interpretationof human characteristics. Therefore, it is calling for new effective exploratory analysis tools.We introduce a latent theme dictionary model for modeling a bunch of process data to identifyco-occurrent event patterns as well as cluster individuals with similar behaviors into sub-groups.An estimation approach based on non-parametric Bayes method is proposed, it performs wellon the simulated datasets. We also apply our method to a real data set, from Traffic item inthe 2012 Programme for International Student Assessment.

## David Hirshberg—Augmented Linear Minimax Estimation

The excellent statistical properties of linear estimators are well known. In the context of estimating linear functionals of regression functions observed with Gaussian noise, it was shown by Donoho in 'Statistical Estimation and Optimal Recovery' that linear estimators are nearly minimax among all estimators over regression functions belonging to a convex class. In this talk, I will discuss how the behavior of these minimax linear estimators can be improved by augmenting them with a regression adjustment. I will show that these augmented minimax linear estimators are semi parametrically efficient with some generality and discuss estimation of several generalizations of the average treatment effect to continuous-valued treatments.

## Lydia Hsu—Variable Selection in Graphical Models

TBD

## Sihan Huang—Pairwise Covariates-adjusted Block Model for Community Detection

One of the most fundamental problems in network study is community detection. The stochastic block model (SBM) is one widely used model for network data with different estimation methods developed with their community detection consistency results unveiled. However, the SBM is restricted by the strong assumption that all nodes in the same community are stochastically equivalent, which may not be suitable for practical applications. We introduce pairwise covariates-adjusted stochastic block model (PCABM), a generalization of SBM

that incorporates pairwise covariate information. In our model, the pairwise covariates can be constructed using any bivariate function of the corresponding covariates of the pair of nodes considered. We study the maximum likelihood estimators of the coefficients for the covariates as well as the community assignments. It is shown that both the coefficient estimates of the covariates and the community assignments are consistent under typical sparsity conditions. Spectral clustering with adjustment (SCWA) is introduced to efficiently solve PCABM. Under certain conditions, we derive the error bound of community estimation under SCWA and show that it is community detection consistent. PCABM compares favorably with the SBM or degree-corrected stochastic block model (DCBM) under a wide range of simulated and real networks when covariate information is accessible.

## Timothy Jones—Aggregated Relational Counts for Citation Networks

Oftentimes networks are prohibitively large to fit a model. The most common approach to remedy this situation is to subsample the network so that the subgraph contains the same statistical information as the original graph. We take a different approach by using a full network motivated model that can be fit to the entire graph using aggregated relational counts. This talk will define aggregated relational counts and derive a model and inference algorithm for this data. We will conclude with an application to a large citation network.

## Peter Lee—Challenges in spike sorting and YASS: Yet Another Spike Sorter

Spike sorting is a critical first step in extracting neural signals from large-scale electrophysiological data. The talk will describe an efficient, reliable pipeline for spike sorting on dense multi-electrode arrays (MEAs), where neural signals appear across many electrodes and spike sorting currently represents a major computational bottleneck. The pipeline is based on an efficient multi-stage "triage-then-cluster-then-pursuit" approach that initially extracts and clusters only clean, high-quality waveforms from the electrophysiological time series by temporarily skipping noisy or "collided" events (representing two neurons firing synchronously). Then, the "triaged" waveforms are then finally recovered with matching-pursuit deconvolution techniques.

## Hok Kan Ling—A proportional intensity model with random effects for process data

We propose a general proportional intensity model with random effects for process data. In an exploratory analysis on process data, a large number of possibly time-varying covariates maybe included. To achieve variable selection for both the fixed effects and the random effects, we impose a penalty term on the log-likelihood. The computation is carried out by the EM algorithm combined with the coordinate descent algorithm. Simulation studies demonstrate that the proposed estimators and estimation algorithm provide an effective recovery of the true structure in both the fixed effects and the random effects. The method is applied to a process data from an item in the problem solving in technology rich environment domain in the Programme for the International Assessment of Adult Competencies.

## Gabriel Loaiza Ganem—Dimensionality Reduction for Point Process Data

In this talk I will present a generative model for point process data. In the model, each point process has a corresponding low-dimensional hidden variable, which is then mapped through a non-linear function to obtain the intensity function (parametrized as a positive spline) of the corresponding point process. Approximate Bayesian inference is performed to recover the posterior distribution of the hidden variables. I will present some preliminary results showing the models performance.

## Florian Stebegg—Robust Pricing and Hedging around the Globe

(joint work with Sebastian Herrmann)

We consider the Robust Pricing Problem for a class of options encompassing American, Asian, Bermudan and European Options in a martingale optimal transport setting for càdlàg processes. We prove strong duality of the pricing and hedging problem and the existence of an optimal pathwise hedge. Our approach provides insight into the structure of primal and dual optimizers which allows us to reduce the problem to a semi-infinite linear program in the case of finitely supported marginals.

## Yixin Wang—Frequentist Consistency of Variational Bayes

(joint work with David Blei)

A key challenge for modernBayesianstatistics is how to perform scalableinferenceof posterior distributions. To address this challenge, variational Bayes (VB) methods have emerged as a popular alternative to the classical Markov chain Monte Carlo MCMC) methods. VB methods tend to be faster while achieving comparable predictive performance. However, there are few theoretical results around VB. In this paper, we establishfrequentistconsistencyand asymptotic normality of VB methods. Specifically, we connect VB methods to point estimates based onvariationalapproximations, calledfrequentist variationalapproximations, and we use the connection to prove a variationalBernsteinVon Mises theorem. The theorem leverages the theoretical characterizations offrequentistvariational approximations to understand asymptotic properties of VB. In summary, we prove that (1) the VB posterior converges to the Kullback-Leibler (KL) minimizer of a normal distribution, centered at the truth and (2) the correspondingvariationalexpectation of the parameter is consistentand asymptotically normal. As applications of the theorem, we derive asymptotic properties of VB posteriors inBayesianmixture models,Bayesiangeneralized linear mixed models, andBayesian stochastic block models. We conduct a simulation study to illustrate these theoretical results.

## Zhi Wang—Improving approximation via adaptive weighting

We propose an adaptive sampling distribution for stochastic gradient descent. By minimizing the trace of the variance-covariance matrix of stochastic updates, the optimum can be approximated in a more aggressive manner. When applied to experiment design or active learning problems, the sampling probability for each candidate is proportional to the square root of the trace of its Fisher information matrix. A subset sampling scheme can reduce the computational burden while retaining the power of weighted sampling. We analyze a real knowledge graph completion example and demonstrate the boost in prediction accuracy under various settings.

## Jing Wu—Modeling Sporadic Event Dynamics with Markov-Modulated Hawkes Processes

Modeling event dynamics is central to many disciplines. In particular, point processes models have been applied to explain patterns seen in event arrival times. Such data often exhibits heterogeneous and sporadic trends, which is challenging to conventional methods. It is reasonable to assume that there exists a hidden state process that drives different event dynamics at different states. In this paper, we propose a Markov Modulated Hawkes Process (MMHP) model and develop corresponding inference algorithms. Numerical experiments using synthetic data and data from an animal behavior study demonstrate that MMHP

with the proposed estimation algorithms consistently recover the true hidden state process in simulations, and separately captures distinct event dynamics with interesting social structure in real data.

## Yuling Yao—Yes, but Did It Work?: Evaluating Variational Inference

While it's always possible to compute a variational approximation to a posterior distribution, it can be difficult to discover problems with this approximation". We propose two diagnostic algorithms to alleviate this problem. The Pareto-smoothed importance sampling (PSIS) diagnostic gives a goodness of fit measurement for joint distributions, while simultaneously improving the error in the estimate. The variational simulation-based calibration (VSBC) assesses the average performance of point estimates.

## Wenda Zhou—Compressibility and generalization in large-scale deep learning

Modern neural networks are highly overparametrized, with capacity to substantially overfit to training data. Nevertheless, these networks often generalize well in practice. It has also been observed that trained networks can often be compressed to much smaller representations. In this talk, we connect these two empirical observations. We present a generalization bound for compressed networks based on the compressed size. Combined with off-the-shelf compression algorithms, we provide the first non-vacuous generalization guarantees for realistic architectures on the ImageNet problem.

# 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 by juveniles 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.