Minghui Yu Memorial Conference

April 22nd 2017
Faculty House

Minghui Yu Memorial Conference
Columbia University Department of Statistics

The Minghui Yu memorial conference

Our annual student conference is held in memory of Minghui Yu, who was a doctoral student at the department of statistics. He passed away in a tragic accident in the spring of 2008.

The conference features talks by the doctoral students of the department and is an opportunity to both remember our friend and colleague and share our research. We would like to thank the department of statistics for its continued support.

Schedule

9:00am—9:25am Breakfast

9:25am—9:30am Opening remarks

9:30am—10:30am Session 1

10:30am—10:45am Morning break

10:45am—12:00pm Session 2

12:00pm—1:00pm Lunch

1:00pm—2:00pm Keynote speaker Persi Diaconis

2:00pm—2:15 pm Break

2:15pm—3:15pm Session 3

3:15pm—3:45pm Afternoon break

3:45pm—5:00pm Session 4

5:45pm Dinner

Location

The conference will be held at the Faculty House (64 Morningside drive) in Garden Room 2.

The dinner will be held at Lido (2168 Frederick Douglass Blvd, New York, NY 10026).

Contact

If you have any question, please do not hesitate to email wz2335@columbia.edu

Persi Diaconis

We are very excited to have Persi Diaconis as our keynote speaker this year. Prof. Diaconis is currently Mary V. Sunseri professor of statistics at Stanford University. As a teenager, Prof. Diaconis was fascinated with magic and magic tricks, which has also inspired some of his mathematical work. Prof. Diaconis started his mathematical education as an undergraduate at City College of New York. He then went on to obtain his Ph.D. from Harvard before becoming Assistant Professor at Stanford. Throughout his career, Prof. Diaconis has held positions at Harvard and Cornell, and has won numerous awards, including the prestigious MacArthur Fellowship and the Rollo-Davidson prize. Prof. Diaconis has made substantial contributions and is a current world expert in many areas of probability, combinatorics and mathematical statistics. He is also widely known for his work on the probability aspects surrounding magic and gambling — including the oft quoted result that seven shuffles will “randomize” a deck of cards.

Keynote talk: THERE'S STUFF TO DO

Statistics is a rapidly changing field. HOWEVER there are lots and lots of concrete problems that remain. This talk is about testing 'randomness' of things like card shuffling schemes (or methods of generating random lottery numbers). There are dozens of instances in which such testing is natural (or even mandatory) but there is very little data (think of a few hundred overhand shuffles of a deck of 52 cards). I want to find easy to implement Bayesian tests which allow incorporation of 'the underlying physics' and common sense. There is a straight-forward way to proceed, I call it Bayesian maximum entropy testing. examples show it seems to do the job. As always, trying to prove things leads to new math and open problems. This is joint work with Guanyang Wang and Sourav Chatterjee.

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.

Student Talks

Session 1—Chair: Tian Zheng

Jing Wu—Latent state model for social interaction events

Social networks, which represent the relationships among actors (such as individuals or companies), have been studied for decades. However, the majority of the existing works model social networks using static or discrete time models. As social networks evolve, a continuous-time model can provide more flexibility in modelling such networks. In this study, we investigate a real data set -- mice fighting data, to study dynamic changes in mice social behaviour. The bursting phenomenon in the dataset motivates us to model the interaction events as inhomogeneous point processes in a relational system. We propose a latent state model to incorporate various interaction behaviour in response to different social context. The result gives a reasonable explanation for the dynamic change in the mice social hierarchy.

Tim 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 for this data. We will conclude with a simulation study that shows using aggregated relational counts can capture the parameters in the full network adequately.

Gabriel Loiza Ganem — Maximum entropy flow networks.

Maximum entropy modeling is a flexible and popular framework for formulating statistical models given partial knowledge. In this work, rather than the traditional method of optimizing over the continuous density directly, we learn a smooth and invertible transformation that maps a simple distribution to the desired maximum entropy distribution. Doing so is nontrivial in that the objective being maximized (entropy) is a function of the density itself. By exploiting recent developments in normalizing flow networks, we cast the maximum entropy problem into a finite-dimensional constrained optimization, and solve the problem by combining stochastic optimization with the augmented Lagrangian method. Simulation results demonstrate the effectiveness of our method, and applications to finance and computer vision show the flexibility and accuracy of using maximum entropy flow networks.

Yixin Wang — Robust probabilistic modeling with Bayesian data reweighting

Probabilistic models analyze data by relying on a set of assumptions. Data that exhibit deviations from these assumptions can undermine inference and prediction quality. Robust models offer protection against mismatch between a model's assumptions and reality. In this talk, we will discuss Bayesian data reweighting, a way to systematically detect and mitigate mismatch of a large class of probabilistic models. This enables robust inference and improves predictive accuracy. We will study different forms of mismatch with reality, ranging from missing latent groups to structure misspecification. A Poisson factorization analysis of the Movielens 1M dataset shows the benefits of this approach in a practical scenario.

Session 2—Chair: Sumit Mukherjee (TBC)

Promit Ghosal — Shock fluctuation of the second class particle in TASEP

We consider totally asymmetric simple exclusion process (TASEP) on $\mathbb{Z}$ where initially all particles in $\mathbb{Z}_{>0}$ are slow and there is a second class particle initially at 0. Slow particles lead to a shock to the system which is guided by the second class particle as the time evolves. Position of the second class particle in TASEP is linked with the competition interface in last passage percolation (LPP) model. The competition interface is introduced and investigated in Ferrari and Leandro (2005), where they found out law of large numbers (LLN) type results. It has been further studied in a recent work Ferrari and Nejjar (2016) where the limiting fluctuation of the competition interface around it LLN limit has been derived under some model assumptions. We further extend its scope to find out the limiting distribution of fluctuation for the second class particle in TASEP when the slow particles on Z<0 generate a shock to the system.

Louis Mittel—The inspection paradox in latent variable models

The Inspection Paradox shows up in many forms and in different problems. It demonstrates how there can be two expectations that are both relevant, yet easily confused with one another. But which expectation are we interested in? I discuss the answer to this question in several examples. Then, I show how a hierarchical latent variable model is a generalization of these examples. This connection helps guide performing correct inference within these models. Several properties of these models are derived.

Lisha Qiu — Numerical methods for stochastic partial differential equation with locally-Lipschitz coefficients

We are interested in convergence properties numerical scheme for SDEs when the assumption of globally Lipschitz continuous coefficient for SDE is extended to locally Lipschitz continuous and no finite time explosion. Convergence in probability and weak convergence are studied. Focusing on the Euler scheme and Milstein scheme, we prove the normalized error processes converge weakly and we also provide the normalizing rates.

Jing Zhang — Modeling time series of counts with shape constraint

In recent years there has been an increasing interest in analysis and modelling of time series of count. Many of the existing estimation methods for count time series models assume the observations follow some known distributions, for example Poisson, Negative Binomial. To relax the distributional assumption, we impose a fairly general shape constraint on the one-parameter exponential family and we propose a semiparametric estimation procedure with a concave component to the observation-driven models studied in Davis and Liu (2012), where the observations are assumed to follow a one-parameter exponential family given the conditional mean process that is modeled as a function of lagged observations. We show that the maximum likelihood estimators of the parameters under the semiparametric framework are consistent. We evaluate the finite sample behavior of the estimators via simulations. Two empirical examples are also provided.

Leo Neufcourt — Expansion of filtration and value of information

Expansion of filtrations is traditionally used to describe insider trading but can also recover various imbalances between market agents. I will explain through a simple example how the additional information of a high-frequency trader can be modeled through an expansion of the filtration with a stochastic process, and show how to derive from the transformation of semimartingales a measure of the advantage given by the additional information, in the form of arbitrage or statistical arbitrage opportunities.

Session 3— Chair: Peter Orbanz

Adji Dieng — The χ-divergence for approximate inference

Variational inference enables Bayesian analysis for complex probabilistic models with massive data sets. It posits a family of approximating distributions and finds the member closest to the posterior. While successful, variational inference methods can run into pathologies; for example, they typically underestimate posterior uncertainty. We propose CHIVI, a complementary algorithm to traditional variational inference. CHIVI is a black box algorithm that minimizes the χ-divergence from the posterior to the family of approximating distributions and provides an upper bound of the model evidence. We studied CHIVI in several scenarios. On Bayesian probit regression and Gaussian process classification it yielded better classification error rates than

EP and classical variational inference (VI). When modeling basketball data with a Cox process, it gave better estimates of posterior uncertainty. Finally, we show how to use the CHIVI upper bound and classical VI lower bound to sandwich estimate the model evidence.

Morgane Austern — Concentration, asymptotic normality and Berry-Esseen rates for group actions and exchangeable structures

Exchangeability is a cornerstone of Bayesian statistics, and two of its statistical consequences are well known: Representation theorems (of de Finetti, Kingman, Aldous-Hoover, etc) and laws of large numbers. We show that stronger properties, namely asymptotic normality of such averages, concentration and Berry-Esseen type of bounds, also follow in a similarly generic fashion. Previously known special cases of this result include Hans Buehlmann's central limit theorem for exchangeable sequences, and the asymptotic normality of certain random graph functionals established by Bickel, Chen and Levina [Ann Statist 39(2011) 38-59]. Additional conditions on the third and fourth moments yield bounds on the rate of convergence that resemble the Berry-Esseen theorem. From exchangeability, the results generalize to other forms of invariance (that is, invariance under actions of other groups than the symmetric group). In these general results, mixing conditions reminiscent of mixing in time series and random fields arises naturally.

Shuaiwen Wang — Which regularizer is optimal for variable selection?

Yuanjun Gao—A structured matrix factorization framework for calcium imaging data analysis and beyond.

Calcium imaging technique enables simultaneous recording of the activities of a large number of neurons. The task of extracting neural activities from the indirect calcium imaging data is nontrivial and involves multiple steps including motion correction, background elimination, region of interest (ROI) detection, calcium deconvolution, etc. Here we introduce the challenges of calcium imaging data and describe a matrix factorization based formulation for ROI detection. The framework decomposes the calcium imaging data (movie data) into a product of structured spatial components (neuron shapes) and temporal components (corresponding neural activities). A fast greedy algorithm is designed to efficiently solve the problem

Session 4—Chair: Michael Sobel

David Hirshberg — Optimizing for efficiency in average treatment effect estimation.

There has been a surge of interest in doubly robust treatment effect estimation in observational studies, driven by a realization that double robustness can be combined with modern machine learning methods to obtain treatment effect estimators that pair semiparametric efficiency with good finite sample performance. A weakness of traditional doubly-robust methods, however, is that they rely on first estimating a propensity model (i.e., the probability that a given subject gets treated given their pretreatment characteristics) and then inverting the estimated propensities; but there is no particularly good reason to believe that an optimally tuned propensity model would also yield optimally tuned inverse-propensity weights. In this paper, we propose an alternative approach to efficient, doubly robust treatment effect estimation based on directly minimizing finite-sample risk bounds via convex optimization tools. Just like classical doubly robust methods, our procedure relies on weighting residuals from a pilot regression; however, our weights are effectively estimated on the inverse-propensity scale. In extensive experiments, we find that our method, functional residual balancing, compares favorably to recently advocated methods such as targeted maximum likelihood estimation or double machine learning.

Feihan Lu — A bayesian hierarchical sparse VAR model for multiple-subject multiple-session resting-state functional connectivity.

Vector autoregressive (VAR) models provide a convenient yet informative framework for learning associations among nodes in graphs. They have been widely applied to a variety of studies, including functional connectivity among brain networks or regions of interest (ROI). However, existing applications have several limitations. First, models that consider connectivity structures for single or multiple subjects within only one imaging session has been developed, and none has been proposed to handle multiple-subject multiple-session imaging data. Second, in many sparse VAR models, the variance-covariance matrix of the innovation process has been explicitly or implicitly assumed to be diagonal, which, despite of its convenience, is not accurate enough for many real practices. In this talk, we present a novel Bayesian hierarchical sparse VAR model for resting-state functional connectivity that generalizes the these limitations: (1). It can be applied to multiple-subject multiple-session data and is capable of simultaneous inference about population-level, subject-level and session-level connectivity over moderately large number of ROI's, along with the variability in connectivity structure among subjects and across sessions; (2). By adopting the doubly adaptive Elastic-net Lasso prior, our model is able to obtain sparse estimates without assuming any special structure of the error covariance matrix. We apply our model to the Human Connectome Project (HCP) data, and we revealed distinct connectivity among 15 ROI's over 7 brain networks.

Kashif Yousuf — Variable screening for high-dimensional time series

This paper analyzes the theoretical properties of Sure Independence Screening (SIS) (Fan and Lv 2008) for high dimensional linear models with dependent and/or heavy tailed covariates and errors. We also introduce a generalized least squares screening (GLSS) procedure which utilizes the serial correlation present in the data. By utilizing this serial correlation when estimating our marginal effects, GLSS is shown to have superior performance in many cases over SIS. Additionally, combining these screening procedures with the adaptive Lasso is analyzed. Dependence is quantified by a functional dependence measure (Wu 2005), and the results rely on the use of Nagaev-type and exponential inequalities for dependent random variables. For both procedures, we derive conditions on the relation between the number of predictors and the sample size, which depend on the moment conditions, strength of dependence of the error and covariate processes, amongst other factors.

Jon Auerbach and Robin Winstanley — No child left behind*

In 2006, the Campaign for Fiscal Equity successfully sued New York State for violating students' constitutional right to a "sound and basic education". The State implemented a formula to disperse education funds based on demographic information. This formula, in tandem with the rapidly changing demography of New York City's diverse neighborhoods, greatly complicates school district planning. We (Robin, Susanna, Swupnil, Tim and Jonathan) predict the growth of students by race, ethnicity and poverty and evaluate the relative influence of spatial and temporal information. We hope this work will help education officials anticipate demographic changes and better budget for future generations of students.

*Our model leaves some children behind.