2016 MINGHUI YU MEMORIAL CONFERENCE

Doctoral Student Body
Department of Statistics
Columbia University
April 23, 2016

We would like to thank the Department of Statistics for their continuous support.

2016 MINGHUI YU MEMORIAL CONFERENCE SCHEDULE

Saturday, April 23 — Social Hall, Union Theological Seminary

9:00	-	9:20	Breakfast
9:20	-	9:30	Opening remarks by Richard Davis, Columbia University
Morning Session I			Chair: Professor Tian Zheng
9:30	-	9:45	Yuting Ma
9:45	-	10:00	Yuanjun Gao
10:00	-	10:15	Swupnil Sahai
10:15	-	10:30	Lu Meng
10:30	-	10:45	Break
Morning Session II			Chair: Professor Marcel Nuts
10:45	-	11:00	Lisha Qiu
11:00	-	11:15	Leo Neufcourt
11:15	-	11:30	Richard Neuberg
11:30	-	11:45	Aditi Dandapani
11:45	-	12:00	Noon break
Morning Session III			Chair: Professor Jose Blanchet
12:00	-	12:15	Phyllis Wan
12:15	-	12:30	Shuaiwen Wang
12:30	-	12:45	Morgane Austern
12:45	-	1:00	Haolei Weng
1:00	-	2:00	Lunch
Keynote Presentation			
2:00	-	3:00	Professor Harrison Zhou, Yale University
3:00	-	3:15	Break
Afternoon Session I			Chair: Professor Jos Zubizarreta
3:15	-	3:30	Maria de los Angeles Resa Juarez
3:30	-	3:45	Susanna Maleka
3:45	-	4:00	Chris Dolan
4:00	-	4:15	Feihan Lu
4:15	-	4:30	Break
Afternoon Session II			Chair: Professor Richard Davis
4:30	-	4:45	Gonzalo Mena
4:45	-	5:00	Jing Zhang

Keynote Presentation

Prof. Harrison Huibin Zhou

Department of Statistics
Yale University

A Review of Some Optimality Results in Network Analysis and Beyond

Abstract: In this talk I will first review some results on graphon estimation and community detection for stochastic block models, then consider extensions to degree corrected stochastic block models, structured matrix completion, and a unified Bayesian posterior contraction framework.

Abstracts

Yuting Ma

Stabilized Sparse Online Learning for Sparse Data

Stochastic gradient descent (SGD) is commonly used for optimization in large-scale machine learning problems. Langford *et al.* (2009) introduce a sparse online learning method to induce sparsity via truncated gradient. With high-dimensional sparse data, however, the method suffers from slow convergence and high variance due to the heterogeneity in feature sparsity. To mitigate this issue, we introduce a stabilized truncated stochastic gradient descent algorithm. We employ a soft-thresholding scheme on the weight vector where the imposed shrinkage is adaptive to the amount of information available in each feature. The variability in the resulted sparse weight vector is further controlled by stability selection integrated with the informative truncation. To facilitate better convergence, we adopt an annealing strategy on the truncation rate, which leads to a balanced trade-off between exploration and exploitation in learning a sparse weight vector. Numerical experiments show that our algorithm compares favorably with the original algorithm in terms of prediction accuracy, achieved sparsity and stability.

Yuanjun Gao

Black Box Variational Inference for State Space Models

Latent variable time-series models are among the most heavily used tools from machine learning and applied statistics. These models have the advantage of learning latent structure both from noisy observations and from the temporal ordering in the data, where it is assumed that meaningful correlation structure exists across time. A few highly-structured models, such as the linear dynamical system with linear-Gaussian observations, have closed-form inference procedures (e.g. the Kalman Filter), but this case is an exception to the general rule that exact posterior inference in more complex generative models is intractable. Consequently, much work in time-series modeling focuses on approximate inference procedures for one particular class of models. Here, we extend recent developments in stochastic variational inference to develop a 'black-box' approximate inference technique for latent variable models with latent dynamical structure.

Swupnil Sahai

Expectation Propagation for Scalable Inference of Hierarchical Bayesian Models

In today's big data age, posterior inference of hierarchical Bayesian models often requires intractable computation times. In particular, HMC's computational costs scale superlinearly in both the number of hierarchical groups and the number of data points per group. Under certain conditions, however, expectation propagation (EP) has been shown to approximate the posterior by learning from subsets of the data at a time. Extending this approach, we show that by splitting hierarchically grouped data into sites and updating the posterior contributions from each site in parallel, EP can significantly cut computational costs while reliably approximating the posterior. We demonstrate the effectiveness of this parallelized EP algorithm on data simulated from a decreasingly parsimonious set of hierarchical models, and conclude by showing how EP's applications to an actual data set in astronomy can help us more efficiently discover radiation from the Big Bang.

Lu Meng

Spectral Filtering for Spatio-temporal Dynamics

Many applications generate spatio-temporal data that exhibit lower-rank smooth movements mixed with higher-rank noises. Separating the signal from the noise is important for us to visualize and understand the lower-rank movements. It is also often the case that the lower rank dynamics have multiple independent components that correspond to different trends or functionality of the system under study. In this presentation, we propose a novel filtering method for identifying lower-rank dynamics and its components embedded in a high dimensional spatio-temporal system, with applications to climate data.

Lisha Qiu

On the Detection of Asset Bubbles

Under the framework of using local martingales to model asset price processes under a risk neutral measure, the detection of asset price bubbles is equivalent to detecting whether or not the price process is a strict local martingale. We model asset price processes with CEV (constant elasticity of variance) processes that have time varying parameters. Some mathematical properties of CEV processes are studied and linked to the severity, or size, of bubbles. The dynamic linear regression method is used to instantaneously detect asset bubbles by estimating time varying parameters of the CEV processes from historical price data. Applications using real equity data are exhibited.

Leo Neufcourt

Insider Trading

Absence of arbitrage opportunities is a fundamental and axiomatic property of mathematical models of financial markets, but it is put into question in a market where some agents possess additional information, canonically modeled by an enlarged filtration. Enlargement of filtrations with random variables has been studied for a long time and is well understood. After recalling the main results in the literature I will show conditions for the absence of arbitrage in the market when the filtration is enlarged with a class of processes particularly suitable to describe the specific information of a high frequency agent.

Richard Neuberg

Predicting Green's Function

Physicists nowadays use machine learning to solve ill-conditioned problems. We show how prior knowledge can be used to improve accuracy by orders of magnitude.

Aditi Dandapani

Initial Expansions of Filtrations and the Strict Local Martingale Property

Strict local martingale arise naturally in applications, most notably in the modeling of financial bubbles. Beginning with a non negative model following a stochastic differential equation with stochastic volatility, we show how a strict local martingale might arise from a true martingale as a result of an enlargement of the underlying filtration. More precisely, we implement a particular type of enlargement, an "initial expansion" of the filtration, for various kinds of stochastic differential equation models, and we provide sufficient conditions such that this expansion can turn a martingale into a strict local martingale. Applications of our work include the modeling and detection of financial bubbles. For example, one might postulate that a bubble arises as a result of the arrival of new information, which we can model via an enlargement of the filtration.

Haolei Weng

Phase Transition and Noise Sensitivity Analysis of Bridge Regression

We study the problem of estimating the coefficient from a linear regression model, under the high dimensional asymptotic regime : we let the dimensionality p go to infinity, while keeping n/p go to a fixed constant (n is the sample size). We consider the popular class of L_q regularized least squares estimators and characterize the almost sure limit of MSE. Based on the expression of the asymptotic MSE, we perform algorithm comparisons (for different q) under noiseless, low noise and large noise regimes. The comparisons will shed light on some peculiar features of phase transition and demonstrate how accurate comparison is possible.

Shuaiwen Wang

Low Noise Analysis of Minimax Estimators

In this project we are interested in characterizing the phase transition for linear regression problem with coefficient vector having general structures. A lower bound is found for the minimum number of observations required to recover the coefficient vector. Some specific cases are studied for the sharpness of the lower bound. Heuristic arguments about whether the bound is sharp in general will be presented.

Morgane Austern

A Central-limit Theorem for Kolmogorov Complexity

Kolmogorov complexity has exhibited promising theoretical results in many estimation problems such as denoising, linear regression, density estimation, etc. However, such theoretical results are overshadowed by the fact that Kolmogorov complexity is not computable. Both the usefulness of the Kolmogorov complexity and its in-computability has motivated researchers to find approximations of this quantity. Hence it has been shown that the average Kolmogorov complexity of a stationary and ergodic source converge to its Shannon entropy. The main question that we would like to address is on the accuracy of this approximation. In particular, we will show that under some regularity conditions on the process, it verifies a central-limit theorem. Our second contribution is concerned with real-valued sequences. One of the quantities that have been observed to play a major role in applications is the Kolmogorov complexity of the quantized process divided by the quantization level. Therefore, we would like to address the following questions: (i) Can we let the quantization step converge to zero with the number of observation? What

would be an appropriate rate? (ii) Can we still obtain a form of central limit theorem for this quantity? We will address those questions.

Phyllis Wan

Asymptotics of Fourier Methods for Testing Independence of Vectors and Time Series

Using the concept of distance covariance, we consider a lag-wise dependence measure of a strictly stationary time seriesthe auto-distance correlation function. For two such sequences we also study the cross-distance correlation function. Assuming strong mixing, we provide asymptotic theory for the empirical auto- and cross-distance correlation functions. We also apply the concept of auto-distance correlation function to the residuals of an autoregressive process and show that limit theory differs from the corresponding one for an iid sequence.

Maria de los Angeles Resa Juarez

Stable Balancing Weights for Marginal Structural Models

Marginal structural models (MSMs) are a very useful tool to estimate the effect of time-dependent treatments in longitudinal studies in the presence of confounders that are also affected by previous treatments. These models appropriately adjust for these covariates by weighing each observation by the inverse of the probability of the observed treatment given the history of observed covariates. However, these probabilities are typically estimated by fitting a model, and the resulting weights can fail to adjust for observed covariates due to model misspecification and also yield very unstable estimates if the predicted probabilities of treatment are very close to zero, which is often the case in practice. To address these problems, instead of modeling the probabilities of treatment, we take a design-based approach and solve a convex optimization problem to directly find the weights of minimum variance that adjust for the covariates across all possible treatment histories. We conduct a simulation study to show the performance of this approach, and find that the proposed weights provide less biased and more precise estimates than other standard methods.

Susanna Maleka

Bayesian Analysis for Cluster Sampling

We develop a Bayesian framework for finite population inference under cluster sampling in a design-based survey context. The two-stage sampling design proceeds by first selecting clusters with probability proportional to cluster sizes and then randomly sampling units from of selected clusters. We incorporate the sampling design into multilevel modeling framework to account for the cluster structure and generalize the inference for non-sampled clusters. The cluster sizes are be treated as covariates. We consider both scenarios when the non-sampled cluster sizes are known and unknown, that is, when the sampling probabilities are known and unknown for the non-sampled units. We estimate the unknown cluster sizes and simultaneously include them as predictors in a flexible hierarchical regression framework with weakly informative prior information. We use simulation studies to evaluate the performance of our procedure and compare it to the classical design-based estimator.

Chris Dolan

Robust Performance of Optimal Switching Problems

Using the Longstaff-Schwarz Algorithm, it is possible to find the solution of an optimal switching problem in the form of state-dependent barriers. It is easy to investigate the performance of this solution when the model is misspecified by simply simulating from an alternate distribution. Using results from the theory of robust optimization, we are able to investigate best/worst case scenarios of model misppecifcation and place upper and lower bounds on the expectation of our value function. We examine case studies taken from the literature on real options pricing in the mining industry

---

Feihan Lu

A Note about Appropriate Background Selection in Disproportionality Analyses of Spontaneous Drug Safety Reports

Post-marketing drug safety surveillance based on large-scale spontaneous reporting system (SRS) databases represents a useful approach for detecting and evaluating potential drug safety issues. Algorithmic methods have emerged to identify signals in SRS databases, including disproportionality analyses, sequential probability ratio tests, and multiple regression. Among those methods, most of the practical experience to date has

been with disproportionality analyses. While the precise operational details of each disproportionality method vary, they all calculate surrogate observed-to-expected ratios in which the reporting experience of each reported drug-event combination is compared to the background reporting experience across all other drugs using an independence model. Disproportionality analyses that compare the reporting experience to one other drug have also been reported. This note shows that the latter comparison using a single drug as background has a masking effect on safety signaling and yields smaller disproportionality metrics more than 90

Ben Reddy

[Title]

[Abstract]

Gonzalo Mena

Extending Spike Sorting for Simultaneous Extra-cellular Electrical Stimulation and Recording

There is a consensus that many new developments in systems neuroscience and neural engineering (e.g. prosthetic devices) will rely in our ability to control neural activity through the simultaneous exogenous stimulation of the neural tissue and the read-out of elicited neural responses. This calls for a methodological shift, as many current methods and technologies were conceived for the mere observation of neurons, but break down in regimes dominated by stimulation. We focus on the simultaneous extra-cellular electrical stimulation and recording, where a proper computational infrastructure for the analysis of neural signals in the large-scale case is still lacking, as spike sorting the method that allows identification of neurons from their electrophysiological spatio-temporal fingerprints cannot handle the corruptions induced by electrical stimuli, since these corruptions the stimulation artifacts can be of much greater magnitude than and overlap with the actual neural activity. In this work we develop such infrastructure, based on a comprehensive account of the variability of the artifact in the spatio-temporal and stimulus dimensions. Specifically, we model the artifact as a Gaussian Process and leverage recent advances in machine learning to enable a fast and scalable implementation. Success is demonstrated by comparison to human-curated ground truth from the primate retina.

Jing Zhang

Semiparametric Estimation for Non-Gaussian Non-minimum Phase ARMA Models

We consider the inference for the parameters of the non-causal and non-invertible ARMA models driven by a non-Gaussian distribution. For such processes, the observations can depend on both the past and future shocks in the system. The non-Gaussianity constraint is necessary to distinguish between causal-invertible and noncausal/noninvertible models. Many of the existing estimation procedures adopt quasi likelihood methods by assuming a non Gaussian noise with distribution fully determined up to an unknown scalar parameter. To relax such distribution restrictions, we borrow ideas from nonparametric density estimation and propose a semiparametric maximum likelihood estimation procedure, in which the noise distribution is projected onto to the space of log-concave measures. We show the maximum likelihood estimators of the conditional likelihood function are consistent. The asymptotic normality of the maximum likelihood estimators is established for AR models

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 and 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 of the Physics Department at Columbia University. One year later, 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.