2015 MINGHUI YU MEMORIAL CONFERENCE

Doctoral Student Body
Department of Statistics
Columbia University
April 18, 2015

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

2015 MINGHUI YU MEMORIAL CONFERENCE SCHEDULE

Saturday, April 18 — Social Hall, Union Theological Seminary

9:00	-	9:50	Breakfast
9:50	-	10:00	Opening remarks by Richard Davis, Columbia University
Morning Session I			Chair: Richard Davis
10:00	-	10:15	Yuanjun Gao
10:15	-	10:30	Zach Shahn
10:30	-	10:45	Yunxiao Chen
10:45	-	11:00	Xiaoou Li
11:05	-	11:20	Break
Morning Session II			Chair: Bodhisattva Sen
11:20	-	11:35	Phyllis Wan
11:35	-	11:50	Susanna Makela
11:50	-	12:05	Jing Zhang
12:05	-	12:20	Rohit Patra
12:20	-	1:20	Lunch

Keynote Presentation

1:20	-	2:20	Professor David Dunson, Duke University

Afternoon Session I			Chair: Lauren Hannah
2:30	-	2:45	Ben Reddy
2:45	-	3:00	Haolei Weng
3:00	-	3:15	Diego Saldana
3:15	-	3:30	Ran He
3:35	-	3:50	Break
Afternoon Session II			Chair: Jose Blanchet
3:50	-	4:05	Lisha Qiu
4:05	-	4:20	Richard Neuberg
4:20	-	4:35	Yang Kang
4:35	-	4:50	Jingjing Zou

Keynote Presentation

Prof. David B. Dunson

The Arts and Sciences Professor of Statistics
Department of Statistical Science
Duke University

Analyzing human brain connection networks

In neuroscience there is increasing interest in relating the structural connection network in white matter tracts (fibers) in the human brain and cognitive traits and neuropsychiatric disorders. There is evidence that the structural network is a more important driver of variability in cognitive traits and disorders than measures of human brain activity (e.g., extracted from fMRI). Recent connectomics pipelines can estimate the brain network based on diffusion tensor imaging and structural MRI. This produces a network-valued random variable for each individual in a study. We develop novel nonparametric Bayes methods for analyzing network-valued data, and for performing inference on the relationship between brain networks and cognitive traits. These methods are provably flexible, reduce dimension adaptively and can be used for formal inferences on group differences adjusting for multiple comparisons automatically. We show dramatic improvements relative to current approaches and illustrate the methods through application to creative reasoning and Alzheimers disease data.

Abstracts

Yunxiao Chen

Latent Variable Model Selection for Binary Response Data via Convex Optimization

We propose a family of models for binary response data that is flexible and jointly models the latent and observed variables in a parsimonious way. This modeling framework combines latent factor model and graphical models (Pearl, 1988; Dawid and Lauritzen, 1993; Lauritzen, 1996) to capture dependence structure not attributable to the latent variables. The rationale is that the latent variables globally drive the dependency among all the observed variables and a sparse graphical model locally characterizes the dependency between the observed variables unexplained by the latent variables. In addition, model estimation is obtained through maximizing a regularized pseudo-likelihood function via a convex optimization algorithm. Using this algorithm, we are able to handle large scale data sets that contain hundreds of observed variables and sample sizes up to millions. Methods are developed to visualize the estimated sparse graphical structure, which helps people understand the dependence structure unexplained by the latent variables. Finally, the model is applied to several data sets from finance, political science, and educational testing.

Yuanjun Gao

Generalized Poisson Linear Dynamical System Model and It's Application to Multineural Recording

Dynamical system models (LDS) have been widely used in analyzing large neural ensemble data. By assuming a low-dimensional latent space that evolves over time, the model captures the inter-neuron correlation and provides a concise representation of the high-dimensional multi-neural data. One limitation of the current LDS model is that the observation model is often assumed to be Poisson, which does not have the flexibility of modeling the under/over-dispersion that is prevalent in neural data. In this paper, we developed generalized Poisson Dynamical System (GPLDS), a generalization of Poisson LDS (PLDS) that models the over/under-dispersion of the neural data. Our model relaxes the Poisson assumption by using larger exponential family for counting data that contains

bernoulli and Poisson distribution as a special case. Variational Bayes Expectation Maximization (VBEM) algorithm is developed for model fitting. Both simulation study and real data analysis show that our model out-performs PLDS in predicting firing rate.

Ran He

A Graphon-based Framework for Modeling Large Networks

Large network, as a form of big data, has received increasing amount of attentions in data science. Analyzing and modeling these data in order to understand the connectivities and dynamics of large networks is important in a wide range of scientific fields. Among popular models, exponential random graph models (ERGMs) have been developed to study these complex networks by directly modeling network structures and features. ERGMs, however, are hard to scale to large networks because maximum likelihood estimation of parameters in these models can be very difficult, due to the unknown normalizing constant. In this talk, I will present a complete computational estimation procedure for estimating ERGMs on large networks. Extensions of this base method in two directions are considered. Inspired by a popular network sampling method, an estimation algorithm using sampled data is proposed, in order to overcome the practical obstacle that the entire network data is hard to obtain and analyze. The base algorithm is also extended to consider the case of complex network structure where nodal attributes exist. Two general frameworks are proposed, one with hierarchical structure, while the other employs similarity measures to incorporate nodal impact.

Yang Kang

Dynamical Models for a Counter-Party Clearing House of an Insurance-Reinsurance Network

We provide a model of dynamic equilibrium settlements in the context of a Counter-Party Clearing House (CPC) for insurance and reinsurance companies. Our main contribution is to show how the widely used Eisenberg-Noe equilibrium settlement mechanism, applied repeatedly over time in a continuous fashion, leads to a suitably defined so-called Skorokhod problem, which is a widely studied process in Stochastic Queueing Networks in Operations Research.

Xiaoou Li

Optimal Adaptive Sequential Design With Application to Crowdsourcing

Commercial crowdsourcing service has gained prominence for obtaining machine learning labels recently. As the crowdsourcing workers are paid for each label they provide, it is desirable to reduce total budget by selecting proper workers adaptively. In this talk, we investigate properties of optimal adaptive sequential designs for worker selection that have minimal Bayes risk. This talk combines several major techniques in statistics and applies them to design optimal crowdsourcing procedure. In particular, we employ techniques in adaptive testing, sequential probability ratio test, stochastic control and empirical Bayes.

Susanna Makela

Instrumental variable strength and sample size in observational studies

Causal inference in observational studies is often accomplished using instrumental variables, a signature method of econometrics. An instrumental variable provides a random push towards acceptance or receipt of treatment and affects the outcome only through this effect on the probability of treatment. Weak instruments are problematic because they are highly sensitive to small biases from unobserved covariates, and standard methods such as two-stage least squares can lead to confidence intervals with inadequate coverage. By adopting an appropriate study design, we can strengthen an instrumental variable in an observational study, but at the price of reduced sample size and potentially the generalizability of causal effects to the larger sample. We illustrate with an application to estimating the effect of incarceration on recidivism using data from the Pennsylvania Commission on Sentencing.

Richard Neuberg

Accounting for Estimation Risk in Pricing under Adverse Selection

Financial product prices are set using mathematical models, which specify a a relationship between a set of inputs, such as a customers credit history and income, and an output, such as an interest rate, using historical data. We examine the challenges involved in estimating model parameters when the counterparty engages in adverse selection: minimizing overall estimation risk, assessing counterparty-specific estimation risk, and determining the necessary compensation for irreducible estimation risk. Pricing regime credit scoring is chosen as an example. Using estimated probabilities of default in a plug-in interest rate

estimator can cause systematic mispricing due to adverse selection as well as bias and variance. We suggest reducing these errors by (i) using an economic model which compensates for estimation risk through bootstrap or asymptotic distributional estimates, and (ii) introducing the use of kernelized logistic regression, a more accurate alternative to commonly used default probability estimators such as logistic regression which also allows to better estimate loan applicant-specific estimation risk. These methods are empirically examined on a panel data set from a German credit bureau, where we study dynamic dependencies such as prior rating migrations and defaults.

Rohit Patra

On Single Index Models with Convex Link

We consider estimation and inference in a single index regression model with an unknown convex link function. In contrast to the standard approach of using kernel methods, we use shape-constrained splines to estimate the convex link function. We develop a method to compute the penalized least squares estimators (PLSEs) of the parametric and the nonparametric components given i.i.d. data. We prove the consistency and find the rates of convergence of the estimators. We establish $n^{-1/2}$-rate of convergence and the asymptotic efficiency of the parametric component under mild assumptions. Our analysis of the parametric component is novel as: the PLSEs lie on the "boundary" of the parameter set (due to the imposed shape constraint), there does not exist a least favorable path through the estimator, and the score of the approximately least favorable path does not satisfy the standard conditions. A finite sample simulation corroborates our asymptotic theory and illustrates the superiority of our procedure over kernel methods, without shape restrictions. The identifiability and existence of the PLSEs are also investigated.

Lisha Qiu

A Local Martingale Model for Detecting Asset Bubbles

Mathematically we define asset bubble as the situation that the price process is a nonnegative strict local martingale. Detecting the time of asset entering and exiting bubble is boiled down into the problem of detecting switching time of price process transferring between martingale and strict local martingale. Under the case that the a risky asset's price being modeled by a Brownian driven SDE(stochastic differential equation), we give conditions on the SDE, to make the assets price process to be a martingale or a strict local martingale under the risk neutral measure and proposed a statistical way to detect asset price bubbles.

Ben Reddy

Bayesian inference in nonexchangeable models of random graphs

I discuss how to perform inference in sequential models of random graphs in which the order that the edges and vertices are introduced carries information. For the case where the order is unobserved, I derive a sampling scheme based on particle MCMC methods and share some preliminary results from experiments on synthetic and real data.

Diego Franco Saldana

How Many Communities Are There?

Stochastic blockmodels and variants thereof are among the most widely used approaches to community detection for social networks and relational data. A stochastic blockmodel partitions the nodes of a network into disjoint sets, called communities. The approach is inherently related to clustering with mixture models; and raises a similar model selection problem for the number of communities. The Bayesian information criterion (BIC) is a popular solution, however, for stochastic blockmodels, the conditional independence assumption given the communities of the endpoints among different edges is usually violated in practice. In this regard, we propose composite likelihood BIC (CL-BIC) to select the number of communities, and we show it is robust against possible misspecifications in the underlying stochastic blockmodel assumptions. We derive the requisite methodology and illustrate the approach using both simulated and real data.

Zach Shahn

Granger Causal Graphical Models for Causal Discovery In Observational Databases With Unobserved Confounders

Recent empirical experiments on large medical insurance claims databases have demonstrated that standard epidemiological methods for detecting causal effects have very low positive predictive value. A main reason is likely the presence of unmeasured confounding. A causal discovery method with higher positive predictive value at the cost of lower sensitivity would be desirable. To this end, we propose an algorithm based on the Granger-causal graphical model framework developed by Michael Eichler. Eichler's insight was that, under assumptions of course, the set of conditional Granger causality relations among

a group of time series variables can sometimes identify associations between pairs of variables from the group as either truly causal or spurious. To evaluate whether a drug causes a potential side effect, for example, our algorithm (1) fits Granger Causal graphs to many small groups of time series variables that contain the drug usage time series and side effect occurrence time series (2) weighs the evidence provided by those graphs that identify the relationship as causal or spurious. This is a work in progress.

Phyllis Wan

Choice of Threshold With a View Towards Inference on Angular Distribution of Regularly Varying Data

Regular variation is often used as the starting point for modeling multivariate heavy-tailed data. A random vector is regularly varying if and only if the radial part R is regularly varying and independent of the angular part Θ as R goes to infinity. To carry out inference for the limiting distribution, a typical strategy is based on the angular components of the data for which the radial parts exceed some threshold. So choosing a large threshold for which the angular and radial parts are nearly independent is an important piece of the inference procedure. In this talk, we would discuss a procedure for choosing the threshold that is based on distance correlation, a measure of independence. We provide some background theory for this procedure and illustrate its performance on both simulated and real data.

Haolei Weng

Variable Selection in Networks

Community detection (clustering for networked data) has been one of the fundamental problems in network analysis. If additional node covariates with the same group structure as the network's are available, the hope is that leveraging covariates can improve the community clustering performance. Both likelihood based and methods of moments (e.g, spectral clustering) have been proposed to accomplish this goal in recent years. In this talk, we consider a more realistic setting in which the covariates include some with different group structures and some without any group structure. Simply putting all of the covariates into clustering algorithm may wash out network's group structure. Motivated by this concern, we propose simple hypothesis testing to screen out the first type of covariates. We show that the testing method is reasonably well as long as network carries strong group signals. After the screening step, we use group-lasso based penalized likelihood method to further screen out the second type of covariates. EM algorithm is used

to do inference. It can be shown that both E-step (approximate belief propagation) and M-step (shrinkage update) can be performed quickly. Finally, we show some preliminary simulation results to demonstrate the improvement of our method.

Jing Zhang

Noncausal vector autoregression with the approximation of log concave distributions

Vector autoregressive models can be considered as the most widely used multivariate time series models. For such processes, the observations may depend on both the past and future shocks in the system. Usually the error distribution is assumed to belong to a fairly general class of elliptical distributions which does not contain gaussian distribution. The exclusion of non-Gaussianity is needed for reasons of identifiability. Then the error distribution is fully determined up to the covariance matrix and a unknown scalar parameter. To release such distribution restrictions, we propose a semi parametric model to learn simultaneously the AR coefficients and the unknown error distribution with the help of the approximation by log concave distributions. We show that the maximum likelihood estimators of the conditional likelihood function are consistent. The estimation procedure is illustrated with a simulation study for a VAR(1) process.

Jingjing Zou

Independent Component Analysis with Under- and Over-complete Features

In the Independent Component Analysis (ICA) models, $X = AS$, where S is the independent components and A is a non-singular transformation matrix. The goal is to recover A and S from the observation X. Samworth and Yuan (2012) proposed a nonparametric maximum likelihood method for undercomplete ICA (with $dim(X) = dim(S)$), in which log-concave projections of empirical distributions of S are used to find the MLE. Here we demonstrate this method with simulation studies and introduce the difficulties and findings when tried to expand the model to overcomplete ICA models ($dim(X) < dim(S)$)

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.