2013 MINGHUI YU MEMORIAL CONFERENCE

Doctoral Student Body
Department of Statistics
Columbia University
April 13, 2013

Thank you to the Department of Statistics and the Graduate Student Advisory Council for their generous support.

2013 MINGHUI YU MEMORIAL CONFERENCE SCHEDULE

Saturday, April 13 — Faculty House, Presidential Room 1

9:30	-	10:00	Breakfast
Morning Session I			Chair: Richard Davis
10:00	-	10:25	Xuan Yang
10:25	-	10:50	Yuwei Zhao
10:50	-	11:15	Xiaoou Li
11:15	-	11:30	Break
Morning Session II			Chair: Peter Orbanz
11:30	-	11:55	Rohit Patra
11:55	-	12:20	Junyi Zhang
12:20	-	1:20	Lunch

Keynote Presentation

1:20	-	2:20	Professor James Berger, Duke University

Afternoon Session I			Chair: Yang Feng
2:30	-	2:55	Stephanie Zhang
2:55	-	3:20	Ying Liu
3:20	-	3:55	Ruixue Fan
3:55	-	4:10	Break
Afternoon Session II			Chair: David Madigan
4:10	-	4:35	Richard Neuberg
4:35	-	5:00	Wei Wang

Keynote Presentation

Prof. James Berger

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

Reproducibility of Science: P-values and Multiplicity

Published scientific findings seem to be increasingly failing efforts at replication. This is undoubtedly due to many sources, including specifics of individual scientific cultures and overall scientific biases such as publication bias. While these will be briefly discussed, the talk will focus on the all-too-common misuse of p-values and failure to properly account for multiplicities as two likely major contributors to the lack of reproducibility. The Bayesian approaches to both testing and multiplicity will be highlighted as possible general solutions to the problem.

Abstracts

Ruixue Fan

A Novel Statistical Approach for Rare Variants Association Studies Incorporating both Marginal and Interaction Effects

Although genome wide association studies have identified many genetic markers associated with human diseases, the variants identified so far explain only a small fraction of disease heritability. More and more evidence suggests that rare variants with much lower minor allele frequencies play a significant role in disease etiology and several burden tests have been proposed to assess the effect of rare variants. One major limitation of these existing methods is that they overlook the interaction effects among rare variants. In this study, we propose a novel statistical procedure that accounts for both marginal and interaction effects of rare variants. The proposed method exhibits higher power in the presence of complicated interaction patterns.

Xiaoou Li

Rare-event Simulation for the Supremum of Gaussian Random Fields: An Old Folk Song Sung to a Faster New Tune

We consider a classic problem concerning the high excursion probabilities of a Gaussian random field f living on a compact set T. We develop efficient computational methods for the tail probabilities $P(\sup_T f(t) > b)$ and the conditional expectations $E(\Gamma(f)|\sup_T f(t) > b)$ as $b \to \infty$. For each ε positive, we present Monte Carlo algorithms that run in *constant* time and compute the interesting quantities with ε relative error. The efficiency results are applicable to a large class of Hölder continuous Gaussian random fields. Besides computations, the proposed change of measure and its analysis techniques have several theoretical and practical indications in the asymptotic analysis of extremes of Gaussian random fields.

Ying Liu

Simulation-efficient Shortest Probability Intervals

Bayesian highest posterior density (HPD) intervals can be estimated directly from simulations via empirical shortest intervals. Unfortunately, these can be noisy (that is, have a high Monte Carlo error). We derive an optimal weighting strategy using bootstrap and quadratic programming to obtain a more computationally stable HPD, or in general, shortest probability interval (Spin). We prove the consistency of our method. Simulation studies on a range of theoretical and real-data examples, some with symmetric and some with asymmetric posterior densities, show that intervals constructed using Spin have better coverage (relative to the posterior distribution) and lower Monte Carlo error than empirical shortest intervals. We implement the new method in an R package (`SPIn`) so it can be routinely used in post-processing of Bayesian simulations.

Richard Neuberg

Predictive Modeling and its Application in Credit Scoring

The differences between explanation and prediction are often not clear. Under the explanatory paradigm, models are optimized to predict best possible, with respect to constraints such as proximity to the data generating process, interpretability and parsimonious modeling. From these constraints it follows that the predictive power of explanatory models is expected to be lower than the power of models optimized for prediction only. This talk provides both an overview of the predictive paradigm, and presents its application in the area of credit scoring. We show that only a predictive model maximizes the return of a credit institution. Using data from a German credit inquiry agency, we demonstrate how such a model can be specified for a finite population of potential credit inquirers. With kernel logistic regression a powerful statistical learning algorithm is applied.

Rohit Patra

Estimation of a Two-component Mixture Model with Applications to Multiple Testing

We consider a two-component mixture model with one known component. We develop methods for estimating the mixing proportion and the other unknown distribution nonparametrically, given i.i.d. data from the mixture model. We use ideas from shape restricted function estimation and develop estimators that are easily implementable and have good finite sample performance. We establish the consistency of our procedures. Completely automated distribution-free finite sample lower confidence bounds are developed for the mixing proportion. The identifiability of the model, and the estimation of the density of the unknown mixing distribution are also addressed. The asymptotic properties of the estimate is also analyzed. We discuss the connection with the problem of multiple testing and compare our procedure with some of the existing methods in that area through simulation studies. We also analyze two data sets, one arising from an application in astronomy and the other from a microarray experiment.

Wei Wang

Challenges with the Use of Cross-validation for Comparing Structured Models

As a simple and compelling approach for estimating out-of-sample predictive error, cross-validation naturally lends itself to the task of model comparison. However, we feel that the legitimacy of cross-validation methods in model comparison is often being taken for granted. In this work, we want to clarify what cross-validation methods are measuring when they are used for model comparison. Using a hierarchical model fit to large survey data with a battery of questions, we show that even though cross-validation might give good estimates of out-of-sample performance, it is not always a sensitive instrument for model comparison. In addition, we emphasize the importance of proper calibration in assisting us interpret the practical importance of the results when conducting model comparison.

Xuan Yang

Evaluating the Efficiency of Markov Chain Monte Carlo From a Large Deviations Point of View

A generic performance measure is proposed to evaluate the efficiency of Markov chain Monte Carlo (MCMC) algorithms. More precisely, the numerical integration problem is considered and the large deviations rate of the probability that the Monte Carlo estimator deviates from the true by a certain distance is used as a measure of efficiency of a particular MCMC algorithm. Numerical methods are proposed for the computation of the rate function based on samples of the renewal cycles of the Markov chain.

Yuwei Zhao

A Fourier Analysis of Extreme Events

The extremogram is an asymptotic correlogram for extreme events constructed from a regularly varying stationary sequence. In this paper, we define a frequency domain analog of the correlogram: a periodogram generated from a suitable sequence of indicator functions of rare events. We derive basic properties of the periodogram such as the asymptotic independence at the Fourier frequencies and use this property to show that weighted versions of the periodogram are consistent estimators of a spectral density derived from the extremogram.

Junyi Zhang

A Step-wise Multiple Testing Procedure for Regression Models

The resting energy expenditure (REE) is crucial to human energy metabolism. One popular REE model was proposed by Elia in 1992. It associates the REE with organs/tissues masses and their metabolic rates. Although this model is widely used, the metabolic rates suggested by Elia have never been tested statistically. The major difficulty of the testing problem is the strong multi-collinearity among organs/tissues masses. This, in turn, makes the least square fit extremely unstable in the sense that some fitted metabolic rates are negative and most confidence intervals are extraordinarily wide. To address the issue of multi-collinearities, we proposed a new multiple testing method called mini-max marginal regression distance (MMRD) step-down procedure. In this short talk, I am going to describe this testing procedure and illustrate it with the REE dataset. The MMRD procedure controls the family-wise error rate (FWER) in a natural way.

Stephanie Zhang

Non-identifiability, equivalence classes, and attribute-specific classification in Q-matrix based Cognitive Diagnosis Models

There has been growing interest in recent years in Q-matrix based cognitive diagnosis models. Parameter estimation and respondent classification under these models may suffer due to identifiability issues. Non-identifiability can be described by a partition separating attribute profiles into groups of those with identical likelihoods. Marginal identifiability concerns the identifiability of individual attributes. Maximum likelihood estimation of the proportion of respondents within each equivalence class is consistent, making possible a new measure of assessment quality reporting the proportion of respondents for whom each individual attribute is marginally identifiable. Arising from this is a new posterior-based classification method adjusting for non-identifiability.

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.