THE UNIVERSITY OF CHICAGO | Stevanovich Center for Financial Mathematics

New Aspects of Statistics, Financial Econometrics, and Data Science

May 10-12, 2018

This event is made possible by the generous philanthropy of University of Chicago Trustee Steve G. Stevanovich, AB '85, MBA '90.

Program

Thursday, May 10

8:00 AM	Registration	
8:30 AM	Breakfast & opening remarks	
9:00-9:50 AM	Yoav Benjamini *Tel Aviv University*	Selective Model Searching
9:50-10:40 AM	Nathan Srebro *Toyota Technological Institute at Chicago*	The Everlasting Database: Validation at a Fair Price
10:40 AM	Break	
11:00-11:50 AM	Kathryn Roeder *Carnegie Mellon University*	Learning from High-dimensional and Noisy Transcriptome Data
12:00 PM	Lunch	
1:30-2:20 PM	Jiashun Jin *Carnegie Mellon University*	Network Analysis by SCORE
2:20-3:10 PM	Matthew Stephens *University of Chicago*	Come Join the Multiple Testing Party!
3:10 PM	Break	
3:40-4:30 PM	Florentina Bunea *Cornell University*	Optimal Estimation of Structured Matrices: Factor Models, Overlapping Clustering and Topic Models
4:30-5:20 PM	Regina Liu *Rutgers University*	Fusion and Individualized Fusion Learning from Diverse Data Sources
5:20 PM	Day 1 concludes	

Friday, May 11

8:30 AM	Breakfast	
9:00-9:50 AM	Raymond Carroll *Texas A&M University*	Semiparametric Analysis of Complex Polygenic Gene-Environment Interactions in Case-Control Studies
9:50-10:40 AM	Jianqing Fan *Princeton University*	Uniform Perturbation Analysis of Eigenspaces and its Applications to Community Detection, Ranking and Beyond

10:40 AM	Break	
11:00-11:50 AM	Xihong Lin *Harvard University*	Test for a Large Number of Composite Null Hypotheses with Application to Mediation Analysis
12:00 PM	Lunch	
1:30-2:20 PM	Francis Diebold *University of Pennsylvania*	Egalitarian LASSO for Combining Economic Forecasts
2:20-3:10 PM	Lan Zhang *University of Illinois at Chicago*	The Five Trolls Under the Bridge: Principal Component Analysis with Asynchronous and Noisy High Frequency Data
3:10 PM	Break	
3:40-4:30 PM	Richard Freeman *Harvard University*	The Rise of China in Global Science: Papers, Citations, and Fields
4:30-5:20 PM	James Evans *University of Chicago*	Centralized Communities More Likely Generate Non-replicable Results
5:30 PM	Reception & posters	
6:30 PM	Day 2 concludes	

Saturday, May 12

8:30 AM	Breakfast	
9:00-9:50 AM	Boaz Nadler *Weizmann Institute of Science*	Unsupervised Ensemble Learning (or How to Make Good Predictions while Knowing Almost Nothing)
9:50-10:40 AM	Dacheng Xiu *University of Chicago*	Empirical Asset Pricing via Machine Learning
10:40 AM	Break	
11:00-11:50 AM	Yingying Li *Hong Kong University of Science and Technology*	Approaching Mean-Variance Efficiency for Large Portfolios
12:00 PM	Lunch	
1:30 PM	Conference ends	

Join our mailing list to receive announcements regarding future conferences!

Abstracts

Yoav Benjamini

Selective Model Searching

Some of the problems facing analysts of treatment-based medical database will be reviewed. The challenge of searching through many possible models, where the dependent variable is not determined pre-analysis, and then inferring on the coefficients of the explanatory variables in the few selected promising models will be discussed. A hierarchical inference approach will be presented and some variations of it compared. The work is motivated and demonstrated by the analysis of a database of Parkinson Disease patients in Tel Aviv Medical Centre, which was done as part of the Health Informatics pillar of the Human Brain Project.

Florentina Bunea

Optimal Estimation of Structured Matrices: Factor Models, Overlapping Clustering and Topic Models

In this talk we introduce a novel estimation method, called LOVE, of the entries and structure of a $p \times K$ loading matrix A in a sparse latent factor model, from n independent observations on X. The unobservable latent factors $Z \in R^K$ are correlated, and their number K is not known prior to estimation. Under assumptions that ensure parameter identifiability, our procedure yields minimax optimal estimators of A, up to logarithmic factors in p. The minimax lower bounds, and the matching upper bounds are valid when K and p are allowed to grow, and be larger than the sample size n. The LOVE procedure scales well with the dimensions of the model, and has overall computational complexity of order p^2. We provide an example of the usage of this model to overlapping clustering, and offer theoretical guarantees for the recovery of the model-based clusters. We also consider the related, but different, problem of estimation in topic models, with an unknown number of topics. If one observes n independent multinomials of dimension p, the topic models postulate a certain factorization of the expectation of the $p \times n$ data matrix. We provide conditions under which the factors are identifiable, and concentrate on the estimation of one of them, called the word-topic matrix, of dimension $p \times K$, with K unknown and growing with n. Using LOVE as background, we develop a new fast algorithm tailored to estimation in these models. We establish minimax lower bounds for the estimation of the word-topic matrix, valid when K and p grow with n, and show that our procedure is minimax-optimal, under minimal conditions.

Raymond J. Carroll

Semiparametric Analysis of Complex Polygenic Gene-Environment Interactions in Case-Control Studies

Many methods have been proposed recently for efficient analysis of case-control studies of gene-environment interactions using a retrospective likelihood framework that exploits the natural assumption of gene-environment independence in the underlying population. We will review some of this literature and discuss some of the fairly astonishing gains in efficiency that are possible for understanding the interactions. However, for polygenic modeling of gene-environment interactions, a topic of increasing scientific interest, applications of retrospective methods have been limited due to a requirement in the literature for parametric modeling of the distribution of the genetic factors, which is difficult because of the complex nature of polygenic data. We propose a fully general, computationally simple, efficient semiparametric method for analysis of case-control studies that allows exploitation of the assumption of gene-environment independence without any further parametric modeling assumptions about the marginal distributions of any of the two sets of factors. The method relies on the key observation that an underlying efficient profile likelihood depends on the distribution of genetic factors only through certain expectation terms that can be evaluated empirically. We develop asymptotic inferential theory for the estimator and evaluate numerical performance using simulation studies. An application of the method is illustrated using a case-control study of breast cancer.

Francis X. Diebold

Egalitarian LASSO for Combining Economic Forecasts

Despite the clear success of forecast combination in many economic environments, several important issues remain incompletely resolved. The issues relate to selection of the set of forecasts to combine, and whether some form of additional regularization (e.g., shrinkage) is desirable. Against this background, and also considering the frequently-found superiority of simple-average combinations, we propose LASSO-based procedures that select and shrink toward equal combining weights. We then provide an empirical assessment of the performance of our "egalitarian LASSO" procedures. The results indicate that simple averages are highly competitive, and that although out-of-sample RMSE improvements on simple averages are possible in principle using egalitarian LASSO methods, they are hard to achieve without endowing the forecaster with information not available ex ante, due to the intrinsic difficulty of small-sample cross validation of LASSO tuning parameters. We therefore propose alternative direct combination procedures, most notably "best average" combination, motivated by the structure of egalitarian LASSO and the lessons learned, which do not require choice of a tuning parameter. Intriguingly, they turn out to outperform simple averages.

James Evans

Centralized Communities More Likely Generate Non-replicable Results

Growing concern that many published results, including those widely agreed upon, may be false are rarely examined against rapidly expanding research production. Exact replications only occur on small scales due to prohibitive expense and limited professional incentive. We introduce a novel, high-throughput replication strategy aligning 51,292 published claims about drug-gene interactions (e.g., Benzo(a)pyrene decreases expression of SLC22A3) with high-throughput experiments performed through the NIH LINCS L1000 program. We propose that the likelihood of a published claim to replicate in future experiments depends in part on how scientific communities are networked in an increasingly collaborative system of "big science". We show (1) that unique claims replicate 19% more frequently than expected, while those widely agreed upon replicate 45% more frequently, manifesting collective correction mechanisms in science. Nevertheless (2) centralized scientific communities perpetuate claims less likely to replicate even if widely agreed upon in the literature and irrespective of biological heterogeneity observed in high-throughput experiments, demonstrating how centralized, overlapping collaborations weaken collective understanding. Decentralized communities in biomedical science involve more independent teams that use more diverse methodologies and draw from more distinctive motivations, generating much more robust, replicable results. Our findings highlight the importance of science policies that foster decentralized collaboration to promote robust scientific advance.

Jianqing Fan

Uniform Perturbation Analysis of Eigenspaces and its Applications to Community Detection, Ranking and Beyond

Spectral methods have been widely used for a large class of challenging problems, ranging from top-K ranking via pairwise comparisons, community detection, factor analysis, among others. Analyses of these spectral methods require super-norm perturbation analysis of top eigenvectors. This allows us to UNIFORMLY approximate elements in eigenvectors by linear functions of the observed random matrix that can be analyzed further. We first establish such an infinity-norm perturbation bound for top eigenvectors and apply the idea to several challenging problems such as top-K ranking, community detections, Z_2-syncronization and matrix completion. We show that the spectral methods are indeed optimal for these problems. We illustrate these methods via simulations. Joint with Emmanuel Abbe, Kaizheng Wang, Yiqiao Zhong and that of Yixin Chen, Cong Ma and Kaizheng Wang.

Richard Freeman

The Rise of China in Global Science: Papers, Citations, and Fields

In less than two decades China has transformed itself from a bit player in global science to giant contributor in virtually every scientific field, as measured by addresses on papers in the largely English language Scopus data base. But China's contribution goes beyond China-addressed papers in Scopus in two substantial ways. Chinese researchers contribute to papers with addresses outside the country as part of the largest diaspora of graduate students, post-docs, visiting and permanent researchers in history. Chinese research has also increased massively in Chinese language journals indexed in China's National Knowledge Infrastructure (CNKI) data base. This paper measures China's increase in physical and natural sciences and separately in social sciences, both in terms of number of papers and citations to papers and examines the role of international collaboration and spread of knowledge through overseas and domestic only Chinese papers. The rise in citations in Chinese papers is related to increased number of Chinese researchers and homophily networks in citations and to improved quality of papers.

Jiashun Jin

Network Analysis by SCORE

We have collected a data set for the networks of statisticians, consisting of titles, authors, abstracts, MSC numbers, keywords, and citation counts of papers published in representative journals in statistics and related fields. In Phase I of our study, the data set covers all published papers from 2003 to 2012 in Annals of Statistics, Biometrika, JASA, and JRSS-B. In Phase II of our study, the data set covers all published papers in 36 journals in statistics and related fields, spanning 40 years. The data sets motivate an array of interesting problems, and for the talk, I will focus on two closely related problems: network community detection, and network membership estimation. We tackle these problems with the recent approach of Spectral Clustering On Ratioed Eigenvectors (SCORE), reveal a surprising simplex structure underlying the networks, and explain why SCORE is the right approach. We use the methods to investigate the Phase I data and report some of the results. We also report some Exploratory Data Analysis (EDA) results including productivity, journal-journal citations, and citation patterns. This part of result is based on Phase II of our data set (ready for use not very long ago).

Yingying Li

Approaching Mean-Variance Efficiency for Large Portfolios

We study the large dimensional Markowitz optimization problem. Given any risk constraint level, we introduce a new approach for estimating the optimal portfolio. The approach relies on a novel unconstrained regression representation of the mean-variance optimization problem, combined with high-dimensional sparse regression methods. Our estimated portfolio, under a mild sparsity assumption, asymptotically achieves mean-variance efficiency and meanwhile effectively controls the risk. To the best of our knowledge, this is the first approach that can achieve these two goals simultaneously for large portfolios. The superior properties of our approach are demonstrated via comprehensive simulation and empirical analysis. Joint work with Mengmeng Ao and Xinghua Zheng.

Xihong Lin

Test for a Large Number of Composite Null Hypotheses with Application to Mediation Analysis

In genome-wide epigenetic studies, it is often of scientific interest in assessing the mediator role of DNA methylation in the causal pathway from an environmental exposure to a clinical outcome. A common approach to mediation analysis consists of fitting two regression models: the mediator model and the outcome model, and then the product of coefficient method was used to estimate the mediation effect and hypothesis testing was performed using Sobel's test. In this paper, we show that the Sobel's method is too conservative for genome-wide epigenetic studies and thus is seriously underpowered to detect mediation effect. We emphasize that the null hypothesis of mediation testing is composite and hence imposes great statistical challenges for assessing mediation effect. In this paper, we propose a divide-aggregate test (DAT) for the detection of mediation effects in genome-wide epigenetic studies by first dividing the composite null parameter space into three disjoint parts and proposing separate testing procedures for each part, and then an overall test is formed by aggregating the statistical evidence from the three parts using the law of total probability with relative proportions of the three parts estimated based on the p-values from the mediator and outcome regression models. We show that this composite testing procedure performs much better than existing methods for genome-wide epigenetic studies where the signals are usually very sparse. A fast Monte Carlo correction is also proposed when DAT indicates slight conservativeness. Simulation studies were conducted to evaluate the type I error rates and powers under various practical settings. An application to the Normative Aging Study (NAS) identified putative DNA methylation CpG sites as mediators in the causal pathway from smoking behavior to lung functions.

Regina Liu

Fusion and Individualized Fusion Learning from Diverse Data Sources

Inferences from different data sources can often be fused together to yield more powerful overall findings than those from individual sources alone. We present a new approach for fusion learning by using the so-called confidence distributions (CD). We further develop the individualized fusion learning, 'iFusion', for drawing efficient individualized inference for a target individual data source or subject by utilizing the leanings from relevant data sources. In essence, iFusion strategically 'borrows strength' from the inferences of relevant individuals to improve the efficiency of the inference of a target individual while retaining its validity and is ideally suited for goal-directed applications such as precision medicine. iFusion is also robust for handling heterogeneity arising from diverse data sources. Computationally, the fusion approach is parallel in nature and scales up well. The performance of the approach is demonstrated by several simulation studies and a project on aviation risk analysis of aircraft landing data.

Boaz Nadler

Unsupervised Ensemble Learning (or How to Make Good Predictions While Knowing Almost Nothing)

In various applications, one is given the advice or predictions of several classifiers of unknown reliability, over multiple questions or queries. This scenario is different from standard supervised learning where classifier accuracy can be assessed from available labeled training or validation data, and raises several questions: given only the predictions of several experts of unknown accuracies, over a large set of unlabeled test data, is it possible to: a) reliably rank them, and b) construct a meta-learner more accurate than any individual experts in the ensemble?
In this talk we'll show that under various independence assumptions between classifier errors, this high dimensional data hides simple low dimensional structures. Exploiting these, we will present simple spectral methods to address the above questions and derive new unsupervised spectral meta-learners. We'll prove these methods are asymptotically consistent when the model assumptions hold and present their empirical success on a variety of unsupervised learning problems.

Kathryn Roeder

Learning from High-dimensional and Noisy Transcriptome Data

Knowing how genes are expressed, how they are co-regulated over development and across different cell types yields insight into how genetic variation translates into risk for complex disease. Here we take on two related statistical challenges in this area: (1) clustering cells based on single cell RNA-sequencing data; and (2) dynamic clustering of genes based on gene expression over developmental periods. To solve these problems, we use global spectral clustering for dynamic networks and semi-soft clustering for single cell gene expression.

Nathan Srebro

The Everlasting Database: Statistical Validity at a Fair Price

The problem of handling adaptivity in data analysis, intentional or not, permeates a variety of fields, including test-set overfitting in ML challenges and the accumulation of invalid scientific discoveries. We propose a mechanism for running a validation service that can answer any arbitrarily long sequence of (potentially adaptive) queries, charging a price for each query and using the proceeds to collect additional samples. Without relying on any declared notion of "users", accounts or adaptivity structure, our pricing mechanism nevertheless ensures analysts making only non-adaptive queries will only pay a very low cost, comparable to the minimal possible cost needed to answer these queries without worrying about adaptivity, while adaptive users bear the cost of answering adaptive queries.

Matthew Stephens

Come Join the Multiple Testing Party!

Multiple testing is often described as a "burden". My goal is to convince you that multiple testing is better viewed as an opportunity, and that instead of laboring under this burden you should be looking for ways to exploit this opportunity. I invite you to multiple testing party.

Dacheng Xiu

Empirical Asset Pricing via Machine Learning

We synthesize the field of machine learning with the canonical problem of empirical asset pricing: Measuring asset risk premia. In the familiar empirical setting of cross section and time series stock return prediction, we perform a comparative analysis of methods in the machine learning repertoire, including generalize linear models, dimension reduction, boosted regression trees, random forests, and neural networks. At the broadest level, we find that machine learning offers an improved description of asset price behavior relative to traditional methods. Our implementation establishes a new standard for accuracy in measuring risk premia summarized by unprecedented high out-of-sample return prediction R2. We identify the best performing methods (trees and neural nets) and trace their predictive gains to allowance of nonlinear predictor interactions that are missed by other methods. Lastly, we find that all methods agree on the same small set of dominant predictive signals that includes variations on momentum, liquidity, and volatility. Improved risk premia measurement through machine learning can simplify the investigation into economic mechanisms of asset pricing and justifies its growing role in innovative financial technologies.

Lan Zhang

The Five Trolls Under the Bridge: Principal Component Analysis with Asynchronous and Noisy High Frequency Data

We develop a principal component analysis (PCA) for high frequency data. As in Northern fairly tales, there are trolls waiting for the explorer. The first three trolls are market microstructure noise, asynchronous sampling times, and edge effects in estimators. To get around these, a robust estimator of the spot covariance matrix is developed based on the Smoothed TSRV (Mykland et al (2017)). The fourth troll is how to pass from estimated time-varying covariance matrix to PCA. Under finite dimensionality, we develop this methodology through the estimation of realized spectral functions. Rates of convergence and central limit theory, as well as an estimator of standard error, are established. The fifth troll is high dimension on top of high frequency, where we also develop PCA. With the help of a new identity concerning the spot principal orthogonal complement, the high-dimensional rates of convergence have been studied after eliminating several strong assumptions in classical PCA. As an application, we show that our first principal component (PC) closely matches but potentially outperforms the S&P 100 market index, while three of the next four PCs are cointegrated with two of the Fama-French non-market factors. From a statistical standpoint, the close match between the first PC and the market index also corroborates this PCA procedure and the underlying S-TSRV matrix. Joint with Dachuan Chen and Per Mykland.