Stevanovich Center
for Financial Mathematics
at the University Of Chicago

5727 South University Avenue
Conference Center (MS112)
Chicago, IL 60637
773-834-8563

October 3-5, 2019

Big Data and Machine Learning in Econometrics, Finance, and Statistics

Abstracts

Conference jointly organized by the University of Chicago Stevanovich Center for Financial Mathematics, the University of North Carolina Rethinc.ML center and the University of Pennsylvania.

It is made possible by the generous philanthropy of University of Chicago Trustee Steve G. Stevanovich.

Andrii Babii	**Machine Learning for Mixed Frequency Data**	In this article, we propose a machine learning method for a high-dimensional time series data that are sampled at different frequencies. We extend the commonly used MIDAS (mixed frequency data sampling) regression model to the high-dimensional setting and study the estimator that builds on the sparse-group LASSO algorithm introduced in Simon et al. (2013). We establish non-asymptotic and asymptotic properties of the estimator with dependent data under mild mixing conditions. Our empirical application to nowcasting the US GPD growth shows that the estimator compares favorably to other alternatives.
Peter Carr	**Using Machine Learning to Predict Realized Variance**	In this paper we formulate a regression problem to predict realized volatility by using option price data and enhance VIX-styled volatility indices' predictability and liquidity. We test algorithms including regularized regression and machine learning methods such as Feedforward Neural Networks on SPX futures and its option data. We noted that in our framework both Ridge regression and FNN can improve volatility indexing with both higher prediction performance and fewer options required. The best performing approach found is to predict the difference between the realized volatility and the VIX-styled index's prediction rather than to predict the realized volatility directly, representing a successful combination of human learning and machine learning. We also discuss suitability of different regression algorithms for volatility indexing and applications of our findings.
Andrew Dai	**Adversarial Methods for Language Understanding and Methods for Multi-modal Data**	Natural language processing with deep learning methods is essential for many fields, including search, recommendation, pricing, advertising, and user modeling. Training deep learning NLP models require large amounts of data and often run into challenging situations where supervised learning is insufficient. For deep learning models such as LSTMs or Seq2Seq, algorithms often require training with unlabeled data to learn broad semantics before supervised training. I will talk about work across semi-supervised learning and supervised learning. I will cover pre-training methods for semi-supervised learning, adversarial training for text classification and finally our recent work on integrating multi-modal data into modelling, including timeseries data and sparse data.

Francis Diebold	**On the Evolution of U.S. Temperature Dynamics**	Climate change is a multidimensional shift. While much research has documented rising mean temperature levels, we also examine range-based measures of daily temperature volatility. Specifically, using data for select U.S. cities over the past half-century, we compare the evolving time series dynamics of the average temperature level, AVG, and the diurnal temperature range, DTR (the difference between the daily maximum and minimum temperatures at a given location). We characterize trend and seasonality in these two series using linear models with time-varying coefficients. These straightforward yet flexible approximations provide evidence of evolving DTR seasonality, stable AVG seasonality, and conditionally Gaussian but heteroskedastic innovations for both DTR and AVG. *Francis X. Diebold, University of Pennsylvania, and Glenn Rudebusch, FRB San Francisco*
Rina Foygel Barber	**Predictive Inference with the Jackknife+**	We introduce the jackknife+, a novel method for constructing predictive confidence intervals that is robust to the distribution of the data. The jackknife+ modifies the well-known jackknife (leave-one-out cross-validation) to account for the variability in the fitted regression function when we subsample the training data. Assuming exchangeable training samples, we prove that the jackknife+ permits rigorous coverage guarantees regardless of the distribution of the data points, for any algorithm that treats the training points symmetrically. Such guarantees are not possible for the original jackknife and we demonstrate examples where the coverage rate may actually vanish. Our theoretical and empirical analysis reveals that the jackknife and jackknife+ intervals achieve nearly exact coverage and have similar lengths whenever the fitting algorithm obeys some form of stability. We also extend to the setting of K-fold cross-validation. Our methods are related to cross-conformal prediction proposed by Vovk [2015] and we discuss connections. *Joint work with Emmanuel Candes, Aaditya Ramdas, and Ryan Tibshirani*

Kay Giesecke	**Towards Explainable AI: Significance Tests for Neural Networks**	Neural networks underpin many of the best-performing AI systems. Their success is largely due to their strong approximation properties, superior predictive performance, and scalability. However, a major caveat is explainability: neural networks are often perceived as black boxes that permit little insight into how predictions are being made. We tackle this issue by developing a pivotal test to assess the statistical significance of the feature variables of a neural network. We propose a gradient-based test statistic and study its asymptotics using nonparametric techniques. The limiting distribution is given by a mixture of chi-square distributions. The tests enable one to discern the impact of individual variables on the prediction of a neural network. The test statistic can be used to rank variables according to their influence. Simulation results illustrate the computational efficiency and the performance of the test. An empirical application to house price valuation highlights the behavior of the test using actual data. *Joint work with Enguerrand Horel, Stanford University*
Jiashun Jin	**Co-authorship and Citation Networks of Statisticians**	We have collected a data set for the networks of statisticians, consisting of titles, authors, abstracts, MSC numbers, keywords, and citation counts of more than 80K papers published in 36 representative journals in statistics and related fields, spanning about 41 years. The data set provides a fertile ground for research in social networks, text mining, and knowledge discovery, and motivates an array of interesting problems in statistics and machine learning. In this talk, we discuss several problems including overall productivity of statisticians, journal ranking, journal clustering, citation patterns, citation prediction, co-authorship network communities, co-authorship network mixed-memberships, dynamic networks, topic estimation, and dynamic topic estimation. Our analysis uses an array of new methods in network analysis, topic estimation, and neural networks.

Tracy Ke	**Optimal Adaptivity of Network Global Testing**	Given a symmetric social network, we are interested in testing whether it has only one community or multiple communities. The desired tests should (a) accommodate severe degree heterogeneity, (b) accommodate mixed-memberships, (c) have a tractable null distribution, (d) adapt automatically to different levels of sparsity, and achieve the optimal phase diagram. How to find such a test is a challenging problem. We propose the Signed Polygon as a class of new tests. Fixing $m \geq 3$, define a score for each for each m-gon in the network, using the centered adjacency matrix. The sum of such scores is then the m-th order Signed Polygon statistic. The Signed Triangle (SgnT) and the Signed Quadrilateral (SgnQ) are special examples of the Signed Polygon. We show that both the SgnT and SgnQ tests satisfy (a)-(d). They especially work well for both very sparse and less sparse networks. Our proposed tests compare favorably with the existing tests. The analysis of the SgnT and SgnQ tests is delicate and extremely tedious. The main reason is that we need a unified proof that covers a wide range of sparsity levels and a wide range of degree heterogeneity. For lower bound theory, we use a phase transition framework, which includes the standard minimax argument, but is more informative.
Bryan Kelly	**The Structure of Economic News**	We propose an approach to measuring the state of the economy via textual analysis of business news. From the full text content of 800,000 Wall Street Journal articles for 1984-2017, we estimate a topic model that summarizes business news as easily interpretable topical themes and quantifies the proportion of news attention allocated to each theme at each point in time. We then use our news attention estimates as inputs into statistical models of numerical economic time series. We demonstrate that these text-based inputs accurately track a wide range of measures of economic activity and that they have incremental forecasting power for macroeconomic outcomes, above and beyond standard numerical predictors. Finally, we use our model to retrieve the news-based narratives that underly shocks to numerical macroeconomic time series.

Po-Ling Loh	**Estimating Location Parameters in Entangled Single-sample Distributions**	We consider the problem of estimating the common mean of independently sampled data, where samples are drawn in a possibly non-identical manner from symmetric, unimodal distributions with a common mean. This generalizes the setting of Gaussian mixture modeling, since the number of distinct mixture components may diverge with the number of observations. We propose an estimator that adapts to the level of heterogeneity in the data, achieving near-optimality in both the i.i.d. setting and some heterogeneous settings, where the fraction of ``low-noise'' points is as small as log(n)/n. Our estimator is a hybrid of the modal interval, shorth, and median estimators from classical statistics; however, the key technical contributions rely on novel empirical process theory results that we derive for independent but non-i.i.d. data. In the multivariate setting, we generalize our theory to mean estimation for mixtures of radially symmetric distributions, and derive minimax lower bounds on the expected error of any estimator that is agnostic to the scales of individual data points. Finally, we describe an extension of our estimators applicable to linear regression. In the multivariate mean estimation and regression settings, we present computationally feasible versions of our estimators that run in time polynomial in the number of data points.
Robert McCulloch	**Searching for Dusty Corners: Understanding the Prediction of the Cross Section of Returns**	We use methods based on ensembles of tree to learn the relationship between predictor variables and the cross-section of expected returns across firms. We focus on tree-based methods because we are able to obtain plausible predictability with minimal assumptions and tuning. Our major goal is to gain some understanding, in a fairly simple way, about the uncovered nonlinear relationship between predictors and expected returns. We use a ``fit-the-fit'' approach in which we look for simplifying structure in our ensemble of tree fit. We first do variable selection by seeking a nonlinear function of a subset of the predictors that approximate the full expectation well. We refit with a reduced number of variables and find that our out-of-sample performance is not diminished. To understand the simplified fit, we fit this fit with a linear model and a generalized additive model and then fit simple trees to the residuals. This gives us a simple description of how the predictions depart from linearity and when the predictions involve interactions amongst the variables. We say that we have discovered ``dusty corners'' in that for most of the predictor space, the linear approximation is representative of the relationship. But, at corners of the space, we can see substantial nonlinearities and interactions involving a relatively small percentage of the observations.

Serena Ng	**Large Dimensional Factor Analysis with Missing Data**	This paper introduces two factor-based imputation procedures that will fill missing values with consistent estimates of the common component. The first method is applicable when the missing data are bunched. The second method is appropriate when the data are missing in a staggered or disorganized manner. Under the strong factor assumption, it is shown that the low rank component can be consistently estimated but there will be at least four convergence rates, and for some entries, re-estimation can accelerate convergence. We provide a complete characterization of the sampling error without requiring regularization or imposing the missing at random assumption as in the machine learning literature. The methodology can be used in a wide range of applications, including estimation of covariances and counterfactuals.
Richard Samworth	**High-dimensional Principal Component Analysis with Heterogeneous Missingness**	We study the problem of high-dimensional Principal Component Analysis (PCA) with missing observations. In simple, homogeneous missingness settings with a noise level of constant order, we show that an existing inverse-probability weighted (IPW) estimator of the leading principal components can (nearly) attain the minimax optimal rate of convergence. However, deeper investigation reveals both that, particularly in more realistic settings where the missingness mechanism is heterogeneous, the empirical performance of the IPW estimator can be unsatisfactory, and moreover that, in the noiseless case, it fails to provide exact recovery of the principal components. Our main contribution, then, is to introduce a new method for high-dimensional PCA, called 'primePCA', that is designed to cope with situations where observations may be missing in a heterogeneous manner. Starting from the IPW estimator, primePCA iteratively projects the observed entries of the data matrix onto the column space of our current estimate to impute the missing entries, and then updates our estimate by computing the leading right singular space of the imputed data matrix. It turns out that the interaction between the heterogeneity of missingness and the low-dimensional structure is crucial in determining the feasibility of the problem. We therefore introduce an incoherence condition on the principal components and prove that in the noiseless case, the error of primePCA converges to zero at a geometric rate when the signal strength is not too small. An important feature of our theoretical guarantees is that they depend on average, as opposed to worst-case, properties of the missingness mechanism.

Erwan Scornet	**A Walk in Random Forests**	The recent and ongoing digital world expansion has led to a growing interest for statistics, as a tool to find patterns in complex data structures, and particularly for turnkey algorithms which do not require specific skills from the user. Such algorithms are quite often designed based on a hunch without any theoretical guarantee. Indeed, the overlay of several simple steps (as in random forests or neural networks) makes the analysis even more arduous. Nonetheless, the theory is vital to give assurance on how algorithms operate therefore preventing their outputs to be misunderstood. In this talk, I will present recent theoretical results on random forests to give insights about how the algorithm works and how aggregation helps to enhance algorithm performance. Special attention will be given to consistency and rate of consistency of several random forests models. In particular, aggregation can turn an inconsistent single tree into a consistent random forests. Moreover, this random forest achieves minimax rate of consistency for twice differentiable functions, which is not true for the single trees that compose the forest.
Hans Skaug	**Integration by Differentiation; Almost a Free Lunch**	The Laplace approximation is a workhorse in statistical computing, providing remarkable accurate integral approximations, also in high dimensions. It involves second order derivatives of the log density, which can automatically be evaluated numerically in arbitrary complex models using a technique called Automatic Differentiation (AD). This algorithm is closely related to Backpropagation, which is a key component of the ongoing deep learning revolution in machine learning. I will report on a decade long experience in developing software based on the idea of combining the Laplace approximation and AD. My examples will be implemented in TMB (https://github.com/kaskr/adcomp). I will also touch upon other AD driven applications of second order derivatives in statistics: saddlepoint approximations, modified profile likelihood, Hamiltonian Monte Carlo, Fisher information matrices. Finally, I will discuss to which extent all of these techniques are applicable and useful in machine learning.

Allan Timmermann	**Do Any Economists Have Superior Forecasting Skills?**	Not really. To answer this question, we develop new methods for testing for superior forecasting skills in settings with arbitrarily many forecasters, outcome variables, and time periods. Our methods allow us to address if any economists had superior forecasting skills for any variables or at any point in time. Answering such questions involves a multiple hypothesis testing problem and requires carefully controlling for the role of "luck" which tends to give rise to false discoveries when a very large number of forecasts are being compared. Exploiting different dimensions of the data, we develop new hypotheses that allow us to identify different types of skills forecasters may possess such as "specialist" skills, limited to one or a few variables with common features, "generalist" skills related to forecasting performance averaged across many variables, or "timing" skills confined to a short window of time or a single period. We apply our new methods to a large set of monthly forecasts of US macroeconomic data and find very little evidence that individual economists can beat a simple equal-weighted average of peer forecasts. *Co-authored with Ritong Qu and Yinchu Zhu.*
Dacheng Xiu	**Predicting Returns with Text Data**	We introduce a new text-mining methodology that extracts sentiment information from news articles to predict asset returns. Unlike more common sentiment scores used for stock return prediction (e.g., those sold by commercial vendors or built with dictionary-based methods), our supervised learning framework constructs a sentiment score that is specifically adapted to the problem of return prediction. Our method proceeds in three steps: 1) isolating a list of sentiment terms via predictive screening, 2) assigning sentiment weights to these words via topic modeling, and 3) aggregating terms into an article-level sentiment score via penalized likelihood. We derive theoretical guarantees on the accuracy of estimates from our model with minimal assumptions. In our empirical analysis, we text-mine one of the most actively monitored streams of news articles in the financial system — the Dow Jones Newswires — and show that our supervised sentiment model excels at extracting return-predictive signals in this context.

Yi Yu	**Dynamic Programming in Change Point Detection in High-Dimensional Autoregression Problems**	In this paper we study the change point detection problems in high-dimensional vector autoregressive models. We assume that the models possess piecewise-constant coefficient matrices and are piecewise stable. We demonstrate the change point localisation consistency of a dynamic programming approach. The model parameters, including the dimensionality, the entry-wise sparsity in the coefficient matrices, the minimal spacing between two consecutive change points and the minimal jump size in terms of the Frobenius norm of the difference of two consecutive coefficient matrices, are allowed to vary with the sample size. We also provide an optional second step in the algorithm to further refining the localisation error rate. Beyond the high-dimensional vector autoregressive models, we also provide a general framework, with the high-dimensional regression problem as a vehicle, unveiling the key ingredients in the consistency of dynamic programming approaches in change point detection in high-dimensional regression and autoregression problems.
Ming Yuan	**Information Based Complexity of High Dimensional Sparse Functions**	We investigate the optimal sample complexity of recovering a general high dimensional smooth and sparse function based on point queries. Our result provides a precise characterization of the potential loss in information when restricting to point queries as opposed to the more general linear queries, as well as the benefit of adaption. In addition, we also developed a general framework for function approximation to mitigate the curse of dimensionality that can also be easily adapted to incorporate further structure such as lower order interactions, leading to sample complexities better than those obtained earlier in the literature.