Stevanovich Center
for Financial Mathematics
at the University Of Chicago

5727 South University Avenue
Chicago, IL 60637
773-834-8563

October 6-8, 2022

Big Data and Machine Learning in

Econometrics, Finance, and Statistics

This conference is made possible by the generous philanthropy of University of Chicago Trustee
Steve G. Stevanovich and the Financial Mathematics Program

Torben Andersen (Northwestern University)

Title: Inference for VIX and Related Option Portfolios

Abstract: In this paper, we develop the first formal inference procedure for risk measures based on option portfolios, such as the VIX. Specifically, for a panel of options with observation errors, we show that the uncertainty about such measures depends critically on the spatial long-run variance (SLRV) of the errors, that is, on their cross-sectional dependence across strikes. Hence, we propose a nonparametric estimator of the SLRV that relies on spatial autocovariance estimates from second-order cross-sectionally differenced observations to extract the parameters of an asymptotically increasing moving average (MA) sequence. Importantly, to accommodate an increasing parameter space for the MA approximation, we propose a new adaptive elastic net minimum distance estimator (AEN-MDE) and establish its asymptotic properties in an infill asymptotic setting -- the mesh of the strike grid for the observed options shrinks asymptotically to zero, while the set of observation times and tenors for the option panel remains fixed. Our novel inference theory, thus, bridges modern high-dimensional methods with nonparametric long-run variance estimation and infill asymptotic limits. A Monte Carlo study shows good finite-sample properties of the developed inference procedure, and an empirical application to S&P 500 index option data reveals that the uncertainty about the VIX and related indices has been declining in recent years. (Joint work with Rasmus Varneskov and Viktor Todorov)

Rina Foygel Barber (University of Chicago)

Title: Conformal prediction beyond exchangeability

Abstract: Conformal prediction is a popular, modern technique for providing valid predictive inference for arbitrary machine learning models. Its validity relies on the assumptions of exchangeability of the data, and symmetry of the given model fitting algorithm as a function of the data. However, exchangeability is often violated when predictive models are deployed in practice. For example, if the data distribution drifts over time, then the data points are no longer exchangeable; moreover, in such settings, we might want to use an algorithm that treats recent observations as more relevant, which would violate the assumption that data points are treated symmetrically. This paper proposes new methodology to deal with both aspects: we use weighted quantiles to introduce robustness against distribution drift, and design a new technique to allow for algorithms that do not treat data points symmetrically, with theoretical results verifying coverage guarantees that are robust to violations of exchangeability.

Matt Dixon (Illinois Tech)

Title: Deep Partial Least Squares for Empirical Asset Pricing

Abstract: We use deep partial least squares (DPLS) to estimate an asset pricing model for individual stock returns that exploits conditioning information in a flexible and dynamic way while attributing excess returns to a small set of statistical risk factors. The novel contribution is to resolve the non-linear factor structure, thus advancing the current paradigm of deep learning in empirical asset pricing which uses linear stochastic discount factors

under an assumption of Gaussian asset returns and factors. This nonlinear factor structure is extracted by using projected least squares to jointly project firm characteristics and asset returns onto a subspace of latent factors and using deep learning to learn the non-linear map from the factor loadings to the asset returns. The result of capturing this non-linear risk factor structure is to characterize anomalies in asset returns by both linear risk factor exposure and interaction effects. Thus the well known ability of deep learning to capture outliers, shed lights on the role of convexity and higher order terms in the latent factor structure on the factor risk premia. On the empirical side, we implement our DPLS factor models and exhibit superior performance to LASSO and plain vanilla deep learning models. Furthermore, our network training times are significantly reduced due to the more parsimonious architecture of DPLS. Specifically, using 3290 assets in the Russell 1000 index over a period of December 1989 to January 2018, we assess our DPLS factor model and generate information ratios that are approximately 1.2x greater than deep learning. DPLS explains variation and pricing errors and identifies the most prominent latent factors and firm characteristics. (Joint work with Nicholas Polson.)

Edgar Dobriban (University of Pennsylvania)

Title: T-Cal: An optimal test for the calibration of predictive models

Abstract: The prediction accuracy of machine learning methods is steadily increasing, but the calibration of their uncertainty predictions poses a significant challenge. Numerous works focus on obtaining well-calibrated predictive models, but less is known about reliably assessing model calibration. This limits our ability to know when algorithms for improving calibration have a real effect, and when their improvements are merely artifacts due to random noise in finite datasets. In this work, we consider detecting mis-calibration of predictive models using a finite validation dataset as a hypothesis testing problem. The null hypothesis is that the predictive model is calibrated, while the alternative hypothesis is that the deviation from calibration is sufficiently large. We find that detecting mis-calibration is only possible when the conditional probabilities of the classes are sufficiently smooth functions of the predictions. When the conditional class probabilities are Hölder continuous, we propose T-Cal, a minimax optimal test for calibration based on a debiased plug-in estimator of the ℓ2-Expected Calibration Error (ECE). We further propose Adaptive T-Cal, a version that is adaptive to unknown smoothness. We verify our theoretical findings with a broad range of experiments, including with several popular deep neural net architectures and several standard post-hoc calibration methods. T-Cal is a practical general-purpose tool, which -- combined with classical tests for discrete-valued predictors -- can be used to test the calibration of essentially any probabilistic classification method.

Jianqing Fan (Princeton University)

Title: How and When are High-Frequency Stock Returns Predictable?

Abstract: This talk presents the studies on the predictability of ultra high-frequency stock returns and durations to relevant price, volume and transactions events, using machine learning methods. We find that, contrary to low frequency and long horizon returns, where predictability is rare and inconsistent, predictability in high frequency returns and durations is large, systematic and pervasive over short horizons. We identify the relevant predictors constructed from trades and quotes data and examine what determines the variation in predictability across different stock's own characteristics and market environments. Next, we compute how the predictability improves with the timeliness of the data on a scale of milliseconds, providing a valuation of each millisecond gained. Finally, we simulate the impact of getting an (imperfect) peek at the incoming order flow, a look ahead

ability that is often attributed to the fastest high frequency traders, in terms of improving the predictability of the following returns and durations.

Eric Ghysels (University of North Carolina at Chapel Hill)

Title: Three common factors

Abstract: Hint: these are not the Fama-French 3 factors and they are not even spanned by the Fama-French 5 factors. More importantly, they feature superior out-of-sample pricing performance compared to standard asset pricing models. What is "common" about these factors? We identify the factor space common between individual stocks and sorted portfolios - neither affected by time-varying betas nor by the sorting characteristics. (Joint work with Elena Andreou, Patrick Gagliardini, and Mirco Rubin.)

Tracy Ke (Harvard University)

Title: Mixed membership estimation for large social networks

Abstract: Given a large network, we assume that there are K perceivable communities and that each node can belong to multiple communities via a mixed membership vector. We are interested in estimating these mixed membership vectors, as they represent the latent social structure of nodes. We propose a spectral method, Mixed-SCORE. It uses the pre-PCA & post-PCA normalizations to simultaneously maximize the signal-to-noise ratio at all entries of eigenvectors and estimates the mixed membership vectors from a simplex geometry in the spectral domain. Under a degree-corrected mixed membership model, we show that Mixed-SCORE is "optimally adaptive": It achieves the optimal rate for many different combinations of network sparsity, degree heterogeneity and signal strength, and the method does not need any prior information of model parameters. For real applications, we apply our method to the MADStat dataset (Ji et al., 2022). It contains attributes of 83K papers published 36 statistics-related journals. We constructed a co-citation network of statisticians and applied Mixed-SCORE to discover a Research Triangle and a Research Map about the academic statistics society. The results provide evidence of the philosophical Research Triangle conjectured by Bradley Efron.

Bryan Kelly (Yale University)

Title: Virtue of Complexity

Abstract: The extant literature predicts market returns with "simple" models that use only a few parameters. Contrary to conventional wisdom, we theoretically prove that simple models severely understate return predictability compared to "complex" models in which the number of parameters exceeds the number of observations. We empirically document the virtue of complexity in US equity market return prediction. Our findings establish the rationale for modeling expected returns through machine learning.

Per Mykland (University of Chicago) and Lan Zhang (University of Illinois at Chicago)

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

Abstract: We develop a principal component analysis (PCA) for high frequency data. As in the fairy 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. 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 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. 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, in the sense of Karl Popper. (Joint work with Dachuan Chen.)

Mark Podolskij (University of Luxembourg)

Title: On Lasso estimation for the drift function in diffusion models

Abstract: In this talk we study the properties of the Lasso estimator of the drift component in the diffusion setting. More specifically, we consider a multivariate parametric diffusion model X observed continuously over the interval [0, T] and investigate drift estimation under sparsity constraints. We allow the dimensions of the model and the parameter space to be large. We obtain an oracle inequality for the Lasso estimator and derive an error bound for the L2-distance using concentration inequalities for linear functionals of diffusion processes. The probabilistic part is based upon elements of empirical processes theory and, in particular, on the chaining method.

Zhao Ren (University of Pittsburgh)

Title: Heteroskedastic Sparse PCA in High Dimensions

Abstract: Principal component analysis (PCA) is one of the most commonly used techniques for dimension reduction and feature extraction. Though it has been well-studied for high-dimensional sparse PCA, little is known when the noise is heteroskedastic, which turns out to be ubiquitous in many scenarios, like biological sequencing data and information network data. We propose an iterative algorithm for sparse PCA in the presence of heteroskedastic noise, which alternatively updates the estimates of the sparse eigenvectors using power method with adaptive thresholdings in one step, and imputes the diagonal values of the sample covariance matrix to reduce the estimation bias due to heteroskedasticity in the other step. Our procedure is computationally fast and provably optimal under the generalized spiked covariance model, assuming the leading eigenvectors are sparse. A comprehensive simulation study demonstrates its robustness and effectiveness under various settings.

Emil Stoltenberg (Oslo School of Business BI)

Title: Regression discontinuity design with right-censored survival data

Abstract: In this paper the regression discontinuity design is adapted to the survival analysis setting with right-censored data. Various causal estimands and estimators for these are introduced. Large-sample theory for the estimators are presented, including confidence intervals that take into account the uncertainty induced by bias correction. As is standard in the causality literature, the models and the theory are embedded in the potential outcomes framework. Two general results pertaining to potential outcomes and the multiplicative hazard model for survival data are presented.

Pragya Sur (Harvard University)

Title: A new central limit theorem for the classical augmented IPW estimator: variance inflation, cross-fit covariance and beyond

Abstract: Estimating the average treatment effect (ATE) is a central problem in causal inference. Modern advances in the field studied estimation and inference for the ATE in high dimensions through a variety of approaches. Doubly robust estimators such as the augmented inverse probability weighting (AIPW) form a popular approach in this context. However, the high-dimensional literature surrounding these estimators relies on sparsity conditions, either on the outcome regression (OR) or the propensity score (PS) model. This talk will introduce a new central limit theorem for the classical AIPW estimator, that applies agnostic to such sparsity-type assumptions. Specifically, we will study properties of the cross-fit version of the estimator under well-specified OR and PS models, and the common modern regime where the number of features and samples are both large and comparable. In this regime, under assumptions on the covariate distribution, our CLT will uncover two crucial phenomena among others: (i) the cross-fit AIPW exhibits a substantial variance inflation that can be precisely quantified in terms of the signal-to-noise ratio and other problem parameters, (ii) the asymptotic covariance between the estimators used while cross-fitting is non-negligible even on the root-n scale. These findings are strikingly different from their classical counterparts, and open a vista of possibilities for studying similar other high-dimensional effects. On the technical front, our work utilizes a novel interplay between three distinct tools—approximate message passing theory, the theory of deterministic equivalents, and the leave-one-out approach. Time permitting, I will outline some of these techniques.

Victor Veitch (University of Chicago)

Title: A Causal View on Invariance and Transportability in Domain Shifts

Abstract: Machine learning methods can be unreliable when deployed in domains that differ from the domains on which they were trained. One intuitive way to address this is to learn a representation of the data that is in some sense "invariant" across the domains, with the hope that models built on this representation will be robust to domain shifts. There are many existing methods following this intuition. Unfortunately, these methods often contradict one another and, empirically, none of them are consistently better than simple vanilla empirical risk minimization (i.e., just ignoring the problem). This begs the question: when, if ever, do we expect each method for invariant representation learning to work? I'll talk about an attack on this problem using a causal view of

domain shifts. I'll also discuss the role of causality in the (closely related) problem of building representations that allow for rapid learning in new domains.

Dacheng Xiu (University of Chicago)

Title: Prediction When Factors are Weak

Abstract: In macroeconomic forecasting, principal component analysis (PCA) has been the most prevalent approach to the recovery of factors, which summarize information in a large set of macro predictors. Nevertheless, the theoretical justification of the PCA-based approach often relies on a convenient and critical assumption that factors are pervasive. To incorporate information from weaker factors, we propose a new prediction procedure based on supervised PCA, which iterates over selection, PCA, and projection. The selection step finds a subset of predictors most correlated with the prediction target, whereas the projection step permits multiple weak factors of distinct strength. We justify our procedure in an asymptotic scheme where both the sample size and the cross-sectional dimension increase at potentially different rates. Our empirical analysis highlights the role of weak factors in predicting inflation.

Paolo Zaffaroni (Imperial College London)

Title: Factor Models for Conditional Asset Pricing

Abstract: This paper develops a methodology for inference on conditional asset pricing models robust to omitted risk factors and to misspecified conditional dynamics. All the features of the asset pricing model, such as risk premia, factors' exposures, factors' variances and covariances, idiosyncratic risk, and number of risk factors, are potentially time- varying. The limiting results hold when the number of assets diverges but the time- series dimension is fixed, possibly very small, applicable to a variety of data frequencies. An extensive empirical application based on individual asset returns data demonstrates the powerfulness of the methodology, allowing to tease out the empirical content of the time-variation elicited by asset pricing theory.

Anru Zhang (Duke University)

Title: Tensor Learning in 2020s: Methodology, Theory, and Applications

Abstract: The analysis of tensor data, i.e., arrays with multiple directions, has become an active research topic in the era of big data. Datasets in the form of tensors arise from a wide range of scientific applications. Tensor methods also provide unique perspectives to many high-dimensional problems, where the observations are not necessarily tensors. Problems in high-dimensional tensors generally possess distinct characteristics that pose great challenges to the data science community. In this talk, we discuss several recent advances in tensor learning and their applications in genomics, computational imaging, and electronic health records. We also illustrate how we develop statistically optimal methods and computationally efficient algorithms that interact with the modern theories of computation, high-dimensional statistics, and non-convex optimization.