Stevanovich Center
for Financial Mathematics
at the University Of Chicago

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

October 3-5, 2024

Big Data and Artificial Intelligence 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.

Yacine Ait-Sahalia (Princeton)

Title: Asset pricing in an economy with changing sentiment and price feedback

Abstract: This paper proposes a continuous-time equilibrium model where the representative agent is subject to a stochastic sentiment. The agent's sentiment responds to the past evolution of prices and can jump. Jumps in sentiment are more likely when sentiment gets more out of line with fundamentals. Feedback takes place from asset prices to sentiment and from sentiment back to asset prices. We show that in equilibrium the agent's sentiment affect prices even though sentiment has no bearing on the asset's fundamentals. Empirically, the equilibrium risk premia and riskfree rate respond to measurable shifts in sentiment in the direction predicted by the model. (Joint work with Patrick Beissner, Patrick Cheridito, & Felix Matthys).

Denis Chetverikov (University of California, Los Angeles)

Title: Estimation of risk premia with many factors

Abstract: *TBA*. (Join work with Nicola Borri, Yukun Liu, & Aleh Tsyvinski).

Francis X. Diebold (University of Pennsylvania)

Title: Machine learning and the yield curve: tree-based macroeconomic regime switching

Abstract: We explore tree-based macroeconomic regime-switching in the context of the dynamic Nelson-Siegel (DNS) yield-curve model. In particular, we customize the tree-growing algorithm to partition macroeconomic variables based on the DNS model's marginal likelihood, thereby identifying regime-shifting patterns in the yield curve. Compared to traditional Markov-switching models, our model offers clear economic interpretation via macroeconomic linkages and ensures computational simplicity. In an empirical application to U.S. Treasury bond yields, we find (1) important yield curve regime switching, and (2) evidence that macroeconomic variables have predictive power for the yield curve when the short rate is high, but not in other regimes, thereby refining the notion of yield curve "macro-spanning". (Joint work with Siyu Bie, Jingyu He, & Junye Li).

Chao Gao (University of Chicago)

Title: Are adaptive robust confidence intervals possible?

Abstract: We consider the problem of confidence interval construction for data with arbitrary contamination. We show that when the contamination proportion is unknown, the length of the optimal robust confidence interval must be exponentially wider.

Fang Han (University of Washington, Seattle)

Title: Chattejee's rank correlation: what is new?

Abstract: In this talk, the speaker will provide an overview of the recent progress made in exploring Sourav Chatterjee's newly introduced rank correlation. The objective is to elaborate on its practical utility and present several new findings pertaining to (a) the asymptotic normality and limiting variance of Chatterjee's rank correlation, (b) its statistical efficiency for testing independence, and (c) the issue of its bootstrap inconsistency. Notably, the presentation will reveal that Chatterjee's rank correlation is root-n consistent, asymptotically normal, but bootstrap inconsistent - a rare phenomenon in the literature.

Ruimeng Hu (University of California, Santa Barbara)

Title: Deep reinforcement learning for games with controlled jump-diffusion dynamics

Abstract: Many real-world multi-party decision-making problems are subject to sudden exogenous events—such as wars, central bank decisions, or global crises like COVID-19—that can cause major shifts in the system, impacting all players simultaneously. These scenarios can be modeled mathematically as games with controlled jump-diffusion dynamics. In this talk, we introduce a computational framework using the actor-critic method in deep reinforcement learning to solve stochastic control problems with jumps. We further extend this algorithm to handle multi-agent games with jumps, utilizing parallel computing to improve computational efficiency. To illustrate the accuracy, efficiency, and robustness of our approach, we provide numerical examples including the Merton problem with jumps, linear quadratic regulators, and the optimal investment game under various conditions.

Bryan Kelly (Yale)

Title: APT or 'AIPT'? The surprising dominance of large factor models

Abstract: We introduce artificial intelligence pricing theory (AIPT). In contrast with the APT's foundational assumption of a low dimensional factor structure in returns, the AIPT conjectures that returns are driven by a large number of factors. We first verify this conjecture empirically and show that nonlinear models with an exorbitant number of factors (many more than the number of training observations or base assets) are far more successful in describing the out-of-sample behavior of asset returns than simpler standard models. We then theoretically characterize the behavior of large factor pricing models, from which we show that the AIPT's "many factors" conjecture faithfully explains our empirical findings, while the APT's "few factors" conjecture is contradicted by the data. (Joint work with Antoine Didisheim, Shikun Ke, & Semyon Malamud).

Mladen Kolar (University of Southern California)

Title: Confidence sets for causal discovery

Abstract: Causal discovery aims to uncover the underlying causal relationships among variables in a multivariate dataset. Traditional methods primarily focus on estimating a single causal model or equivalence class, often neglecting the uncertainty inherent in the causal ordering process. In this talk, we present a novel framework for constructing confidence sets for causal orderings within the context of structural equation models (SEMs). Our approach leverages a residual bootstrap procedure to test the goodness-of-fit of causal orderings, providing a statistically rigorous method for quantifying uncertainty in causal discovery. We establish the asymptotic validity of the proposed confidence sets, which can be used to infer sub/supersets of ancestral relationships and to construct confidence intervals for causal effects that incorporate model uncertainty. This framework is particularly valuable when the signal strength is weak or when key modeling assumptions might be violated, offering a robust tool for analysts to avoid overconfidence in specific causal models. The methodology is computationally efficient, suitable for medium-sized problems, and maintains theoretical guarantees even as the number of variables increases. We illustrate the practical implications of our approach through an analysis of daily stock returns for industry portfolios, highlighting how different causal orderings can lead to significantly different conclusions, and how our confidence sets can inform more reliable causal inferences. This work represents a significant advancement in causal discovery, offering a new dimension of uncertainty quantification that enhances the robustness and reliability of causal inference in complex systems. (Joint work with Sam Wang, & Mathias Drton).

Zongming Ma (Yale)

Title: Multimodal data integration and cross-modal querying via orchestrated approximate message passing

Abstract: The need for multimodal data integration arises naturally when multiple complementary sets of features are measured on the same sample. Under a dependent multifactor model, we develop a fully data-driven orchestrated approximate message passing algorithm for integrating information across these feature sets to achieve statistically optimal signal recovery. In practice, these reference data sets are often queried later by new subjects that are only partially observed. Leveraging on asymptotic normality of estimates generated by our data integration method, we further develop an asymptotically valid prediction set for the latent representation of any such query subject. We demonstrate the prowess of both the data integration and the prediction set construction algorithms on a tri-modal single-cell dataset.

Theodor Misiakiewicz (Yale)

Title: Deterministic equivalents and scaling laws for random feature regression

Abstract: In this talk, we revisit random feature ridge regression (RFRR), a model that has recently gained renewed interest for investigating puzzling phenomena in deep learning—such as double descent, benign overfitting, and scaling laws. Our main contribution is a general deterministic equivalent for the test error of RFRR.Specifically, under a certain concentration property, we show that the test error is well approximated by a closed-form expression that only depends on the feature map eigenvalues. Notably, our approximation guarantee is non-asymptotic, multiplicative, and independent of the feature map dimension—allowing for infinite-dimensional features. This deterministic equivalent can be used to precisely capture the above phenomenology in RFRR. As an example, we derive sharp excess error rates under standard power-law assumptions of the spectrum and target decay. In particular, we tightly characterize the optimal parametrization achieving minimax rate. (Joint work with Basil Saeed, Leonardo Defilippis, & Bruno Loureiro).

Andrea Montanari (Stanford)

Title: Overparametrization in machine learning: insights from linear models

Abstract: Deep learning models are often trained in a regime that is forbidden by classical statistical learning theory. The model complexity can be larger than the sample size and the train error does not concentrate around the test error. In fact, the model complexity can be so large that the network interpolates noisy training data. Despite this, it behaves well on fresh test data, a phenomenon that has been dubbed 'benign overfitting.' I will review recent progress towards a precise quantitative understanding of this phenomenon in linear models and kernel regression. In particular, I will present a recent characterization of ridge regression in Hilbert spaces which provides a unified understanding on several earlier results. (Joint work with Chen Cheng).

Per Mykland (Massachusetts Institute of Technology) & **Lan Zhang** (University of Illinois Chicago)

Title: Estimating the volatility of drift

Abstract: High frequency data (HFD) is an important source of information about financial markets. Recent years have seen the development of HFD-based estimators of volatility, covariance matrices, regression, principal component analysis, asymptotic variances, etc, with corresponding financial applications. Until now, however, there has been no corresponding estimator of the drift in prices. Drift is usually estimated using long time periods (months, years, and up), with little need for HFD. This paper approaches as in-between question: Can we estimate the volatility of drift? This has traditionally been seen as equally difficult as drift itself. In this paper, however, we show that the volatility of drift can be reasonably estimated in periods of a month, using high frequency considerations. The paper provides both a theory (consistency, central limit theorem), and also carries out an empirical analysis. The latter shows that the volatility of drift behaves differently from the ordinary volatility of prices.

Whitney Newey (Massachusetts Institute of Technology)

Title: Automatic debiased machine learning via Riesz regressions

Abstract: A variety of interesting parameters may depend on high dimensional regressions. Machine learning can be used to estimate such parameters. However estimators based on machine learners can be severely biased by regularization and/or model selection. Debiased machine learning uses Neyman orthogonal estimating equations to reduce such biases. Debiased machine learning generally requires estimation of unknown Riesz representers. A primary innovation of this paper is to provide Riesz regression estimators of Riesz representers that depend on the parameter of interest, rather than explicit formulae, and that can employ any machine learner, including neural nets and random forests. End-to-end algorithms emerge where the researcher chooses the parameter of interest and the machine learner and the debiasing follows automatically. Another innovation here is debiased machine learners of parameters depending on generalized regressions, including high-dimensional generalized linear models. An empirical example of automatic debiased machine learning using neural nets is given. We find in Monte Carlo examples that automatic debiasing sometimes performs better than debiasing via inverse propensity scores and never worse. Finite sample mean square error bounds for Riesz regression estimators and asymptotic theory are also given. (Joint work with Victor Chernozhukov, Víctor Quintas-Martínez, & Vasilis Syrgkanis).

Bodhi Sen (Columbia University)

Title: Multivariate, heteroscedastic empirical bayes via nonparametric maximum likelihood

Abstract: n this talk I will describe the nonparametric empirical Bayes (EB) methodology in the (Gaussian) signal plus noise model with multivariate, heteroscedastic errors. This model arises in many large-scale denoising problems (e.g., in astronomy). We consider the nonparametric maximum likelihood estimator (NPMLE) in this setting. We study the characterization, uniqueness, and computation of the NPMLE which estimates the unknown (arbitrary) prior by solving an infinite-dimensional convex optimization problem. The EB posterior means based on the NPMLE have low regret, meaning they closely target the oracle posterior means one would compute with the true prior in hand. We demonstrate the adaptive and near-optimal properties of the NPMLE for density estimation, denoising and deconvolution. (Joint work with Jake Soloff, & Adityanand Guntuboyina).

Katja Smetanina (University of Chicago)

Title: Perceived shocks and impulse responses

Abstract: This paper shows that the information present in many datasets on expec- tations can be leveraged in a new way to infer quantities of key interest in economics: shocks and the associated impulse responses, as perceived by the agents forming the expectations. The information required is a panel of fore- cast revisions of one variable across a term structure of forecast horizons and over time. The idea is to fit a time-varying factor model, which recovers empirical measures of the latent shocks (the factors) and the associated impulse responses (the loadings) at every point in time. Our nonparametric method relies on weak assumptions and deals with the small-sample nature of these data. An application to historical consensus inflation expectations reveals: 1) a single perceived shock that is highly correlated with inflation surprises; 2) a time-varying shape of the impulse response functions that implies a secular de- crease in the perceived persistence of the shock (making long-run expectations more “anchored" over time). (Joint work with Raffaella Giacomini, & Jason Lu).

Pragya Sur (Harvard)

Title: Generalization error of min-norm interpolators in transfer learning

Abstract: Min-norm interpolators naturally emerge as implicit regularized limits of modern machine learning algorithms. Recently, their out-of-distribution risk was studied when test samples are unavailable during training. However, in many applications, a limited amount of test data is typically available during training. The properties of min-norm interpolation in this setting are not well understood. In this talk, I will present a characterization of the generalization error of pooled min-L2-norm interpolation under covariate and model shifts. Our results have several implications. First under model shift, we derive precise thresholds for when transfer learning helps versus hurts. Specifically, adding source data consistently hurts prediction when the signal-to-noise ratio of the target data is low. However, for higher signal-to-noise ratios, transfer learning helps as long as the shift-to-signal ratio lies below a threshold that I will define. Our results also provide a quantification of the optimal number of target samples necessary for minimizing the generalization error. On the other hand, our results uncover that under covariate shift, heterogeneity between domains improves prediction when the source sample size is small relative to the dimension. This makes a case for the more overparameterized the setting, the more beneficial it is to incorporate heterogeneity during learning. (Joint work with Yanke Song and Sohom Bhattacharya).

Chenhao Tan (University of Chicago)

Title: Towards human-centered AI: predicting fatigue and generating hypothesis with LLMs

Abstract: Human-centered AI advocates the shift from emulating humans to empowering people so that AI can benefit humanity. In this talk, I discuss two directions on using LLMs to address challenging tasks for humans. First, I show that LLMs can be used to predict physician fatigue from clinical notes and reveal hidden racial biases: physicians appear more fatigued when seeing Hispanic and Black patients than White patients. Second, I present a recent work on generating novel hypotheses based on observed data. Our algorithm is able to enable an interpretable hypothesis-based classifier that makes accurate predictions. Moreover, the generated hypotheses not only corroborate human-verified theories but also uncover new insights for the tasks. I will conclude with some exciting future directions.

Dacheng Xiu (University of Chicago)

Title: On the theory of autoencoders

Abstract: Autoencoders are pivotal in unsupervised machine learning, widely employed for dimension reduction, feature learning, and signal denoising. This study provides non-asymptotic guarantees for deep autoencoders within a nonlinear factor model framework. We demonstrate that deep autoencoders can effectively retrieve common components from model inputs. The associated error comprises a component that diminishes with increasing dimensionality---akin to the `blessings of dimensionality' observed in linear factor models---and another component that vanishes with an increasing sample size at the optimal nonparametric regression rate, as if the factors were directly observed. Furthermore, we show that the extracted factors converge to the true latent factors, albeit through a functional transformation. We conclude by showcasing three economic applications of autoencoders. (Joint work with Zhouyu Shen).

Wenxin Zhou (University of Illinois Chicago)

Title: Nonparametric expected shortfall regression with tail-robustness

Abstract: Expected Shortfall (ES), also known as superquantile or Conditional Value-at-Risk, has been recognized as an important measure in risk analysis and stochastic optimization. In this talk, we consider a joint regression framework that simultaneously models the conditional quantile and ES of a response variable given a set of covariates, for which the state-of-the-art approach is based on minimizing a joint loss function that is non-differentiable and non-convex. Motivated by the idea of using orthogonal scores to reduce sensitivity to nuisance parameters, we study a two-step framework for fitting joint quantile and ES regression models nonparametrically over RKHSs and using deep neural networks. We establish a non-asymptotic theory for the proposed estimators, carefully characterizing the impact of quantile estimation without relying on sample splitting. For ES kernel ridge regression, we further propose a fast inference method to construct pointwise confidence bands. For ES DNN regression, we introduce a Huberized estimator that is robust against heavy tails in the response distribution. A Python package, quantes (https://pypi.org/project/quantes/), has been developed to implement various ES regression methods.