Stevanovich Center
for Financial Mathematics
at the University Of Chicago

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

October 5-7, 2023

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

Chao Gao (University of Chicago)

Title: Computational lower bounds for graphon estimation via low-degree polynomials

Abstract: Graphon estimation has been one of the most fundamental problems in network analysis and has received considerable attention in the past decade. From the statistical perspective, the minimax error rate of graphon estimation has been established by Gao et al (2015) for both stochastic block model (SBM) and nonparametric graphon estimation. The statistical optimal estimators are based on constrained least squares and have computational complexity exponential in the dimension. From the computational perspective, the best-known polynomial-time estimator is based on universal singular value thresholding (USVT), but it can only achieve a much slower estimation error rate than the minimax one. It is natural to wonder if such a gap is essential. The computational optimality of the USVT or the existence of a computational barrier in graphon estimation has been a long-standing open problem. In this work, we take the first step towards it and provide rigorous evidence for the computational barrier in graphon estimation via low-degree polynomials. Specifically, in both SBM and nonparametric graphon estimation, we show that for low-degree polynomial estimators, their estimation error rates cannot be significantly better than that of the USVT under a wide range of parameter regimes. Our results are proved based on the recent development of low-degree polynomials by Schramm and Wein (2022), while we overcome a few key challenges in applying it to the general graphon estimation problem. By leveraging our main results, we also provide a computational lower bound on the clustering error for community detection in SBM with a growing number of communities and this yields a new piece of evidence for the conjectured Kesten-Stigum threshold for efficient community recovery.

Song Mei (University of California, Berkeley)

Title: Revisiting neural network approximation theory in the age of generative AI

Abstract: Textbooks on deep learning theory primarily perceive neural networks as universal function approximators. While this classical viewpoint is fundamental, it inadequately explains the impressive capabilities of modern generative AI models like GPTs and diffusion models. This talk puts forth a refined perspective: neural networks often serve as algorithm approximators, going beyond mere function approximation. I will explain how this refined perspective offers a deeper insight into the success of GPTs and diffusion models.

Snigdha Panigrahi (University of Michigan)

Title: Selective Inference with Randomized Group LASSO estimators

Abstract: Our work is motivated by the need for inference after regularized estimation with high dimensional datasets that contain grouped covariates. As an example, consider applying a logistic Group LASSO to a dataset with a binary outcome and categorical predictors. How do we conduct selective inference in the estimated sparse model? This problem is challenging due to two reasons: (1) existing approaches for a polyhedral selection method do not apply to the Group LASSO because there is no easy description of the selection event; (2) our data is no longer normal. To solve this problem, we construct an asymptotic selective likelihood that uses extra randomization to obtain an easy to describe selection event. Our new approach provides selective inference

using randomized Group LASSO estimators in likelihood models including generalized linear models, and in other general forms of estimation, such as quasi-likelihood estimation to include possible overdispersion, for example. (Joint work with Yiling Huang, Sarah Pirenne, and Gerda Claeskens.)

Elynn Chen (New York University)

Title: Reinforcement Learning in Latent Heterogeneous Environment

Abstract: Reinforcement Learning holds great promise for data-driven decision-making in various social contexts, including healthcare, education, and business. However, classical methods that focus on the mean of the total return may yield misleading results when dealing with heterogeneous populations typically found in large-scale datasets. To address this issue, we introduce the K-Heterogeneous Markov Decision Process, a framework designed to handle sequential decision problems with latent population heterogeneity. Within this framework, we propose auto-clustered policy evaluation for estimating the value of a given policy and auto-clustered policy iteration for estimating the optimal policy within a parametric policy class. Our auto-clustered algorithms can automatically identify homogeneous subpopulations while simultaneously estimating the action value function and the optimal policy for each subgroup. We establish convergence rates and construct confidence intervals for the estimators. Simulation results support our theoretical findings, and an empirical study conducted on a real medical dataset confirms the presence of value heterogeneity and validates the advantages of our novel approach.

Feng Ruan (Northwestern University)

Title: Kernel Learning “Automatically” Delivers Exactly Low Rank Solutions

Abstract: We consider kernels of the form $(x,x') \mapsto \phi(\norm{x-x'}^2_\Sigma)$ parametrized by Σ. For such kernels, we study a variant of the kernel ridge regression problem which simultaneously optimizes both the prediction function and the reproducing kernel Hilbert space (the latter by optimizing Σ). Assuming that a low-dimensional projection of X is predictive of the response and assuming that covariates have nonzero predictive power, we find that the finite sample kernel learning objective has a global minimizer whose Σ matrix is exactly low rank (with high probability). This phenomenon is interesting because the low rank solution is achieved without using explicit regularization, e.g., nuclear norm penalization. Our theory attributes this low-rankness property in finite samples to the sharpness property of the population objective—a form of structure that’s alike the nuclear norm penalties—at the population minimizers.

Pragya Sur (Harvard University)

Title: Spectrum-Aware Adjustment: A New debiasing paradigm with applications to principal component regression

Abstract: Debiasing methodologies have emerged as a solid toolbox for producing inference in high-dimensional problems. Since its original introduction, the methodology witnessed a major upheaval with the introduction of debiasing with degrees-of-freedom adjustment. Despite overcoming limitations with initial debiasing techniques, this updated method suffers a key issue—the method relies on Gaussian designs and independent, identically

distributed samples. In this talk, we will break this barrier via a novel debiasing formula that exploits the spectrum of the sample covariance matrix. Our formula applies for a broad class of non-Gaussian designs, including some heavy-tailed ones, as well as certain dependent data settings. Our correction term differs significantly from the one proposed in prior work but recovers the Gaussian formula as a special case. We will demonstrate the utility of our approach with regard to several statistical problems such as debiasing the principal component regression estimator, hypothesis testing, confidence interval construction, etc. (Joint work with Yufan Li.)

Linjun Zhang (Rutgers University)

Title: Fair conformal prediction and risk control

Abstract: Multi-calibration is a powerful and evolving concept originating in the field of algorithmic fairness. For a predictor f that estimates the outcome y given covariates x, and for a function class C, multi-calibration requires that the predictor $f(x)$ and outcome y are indistinguishable under the class of auditors in C. Fairness is captured by incorporating demographic subgroups into the class of functions C. Recent work has shown that, by enriching the class C to incorporate appropriate propensity re-weighting functions, multi-calibration also yields target-independent learning, wherein a model trained on a source domain performs well on unseen, future target domains(approximately) captured by the re-weightings. The multi-calibration notion is extended, and the power of an enriched class of mappings is explored. HappyMap, a generalization of multi-calibration, is proposed, which yields a wide range of new applications, including a new fairness notion for uncertainty quantification (conformal prediction), a novel technique for conformal prediction under covariate shift, and a different approach for fair risk control, while also yielding a unified understanding of several existing seemingly disparate algorithmic fairness notions and target-independent learning approaches. A single HappyMap meta-algorithm is given that captures all these results, together with a sufficiency condition for its success.

Heather Battey (Imperial College London)

Title: Inducement of population sparsity

Abstract: The work on parameter orthogonalisation by Cox and Reid (1987) is presented as inducement of population-level sparsity. The latter is taken as a unifying theme for the talk, in which sparsity-inducing parametrisations or data transformations are sought. Three recent examples are framed in this light: sparse parametrisations of covariance matrices; construction of factorisable transformations for the elimination of nuisance parameters; and inference in high-dimensional regression. The solution strategy for the problem of exact or approximate sparsity inducement appears to be context specific and may entail, for instance, solving one or more partial differential equation, or specifying a parametrised path through transformation or parametrisation space.

Marc Hallin (Université Libre de Bruxelles)

Title: Forecasting value-at-risk and expected shortfall in large portfolios: A general dynamic factor model approach

Abstract: Beyond their importance from the regulatory policy point of view, Value-at-Risk (VaR) and Expected

Shortfall (ES) play an important role in risk management, portfolio allocation, capital level requirements, trading systems, and hedging strategies. However, due to the curse of dimensionality, their accurate estimation and forecast in large portfolios is quite a challenge. To tackle this problem, two procedures are proposed. The first one is based on a filtered historical simulation method in which high-dimensional conditional covariance matrices are estimated via a general dynamic factor model with infinite-dimensional factor space and conditionally heteroscedastic factors; the second one is based on a residual-based bootstrap scheme. The two procedures are applied to a panel with concentration ratio close to one. Backtesting and scoring results indicate that both VaR and ES are accurately estimated under both methods, which both outperform the existing alternatives. (Joint work with Carlos Trucíos.)

Nour Meddahi (Toulouse School of Economics)

Title: Non-Linear Time Series Models and Machine Learning

Abstract: We recently observed the irruption and rapid development of maching learning (ML) methods in econometrics and statistics, especially for forecasting purposes. For instance, ML methods have been recently used in several studies for forecasting economic and financial variables like assets returns (Gu, Kelly, and Xiu, 2020), stock and bond returns (Bianchi, Buchner, and Tamoni, 2021), volatility (Patton and Simsek, 2023), inflation (Medeiros, Vasconcelos, Veiga, and Zilberman, 2021), and macroeconomic variables (Goulet Coulombe, 2021; Goulet Coulombe, Leroux, Stevanovic, and Surprenant, 2022). An important common conclusion of these studies is that ML methods are successful in forecasting because they account for non-linearities that popular time series models do not. The first goal of the paper is to highlight the non-linearities that ML methods capture and connect them with traditional non-linear time series modeling. The second goal of the paper is to modify some traditional non-linear time series model by including insights from the ML literature. Applications to the Euro-US dollar exchange rate and the SP500 index are provided. (Joint work with Christian Gourieroux, and Serge Nyawa.)

Torben Andersen (Northwestern University)

Title: The Factor Structure of Systematic Jump Risk

Abstract: This paper develops a test for deciding whether latent systematic jumps in a large cross-section of asset prices obey linear factor structure of low dimension. The test is based on a panel of high-frequency return observations for a cross-section of assets, with both the sampling frequency and the size of the cross-section increasing asymptotically, while the time span of the data set remains fixed. The latent systematic jumps are identified in a nonparametric way and the test statistic is based on evaluating the statistical significance of the smallest eigenvalues of the outer product of the matrix of high-frequency increments that are identified to contain systematic jumps. The limit behavior of the test is highly nonstandard, with systematic diffusive risks as well as idiosyncratic diffusive and jump risks in asset prices all contributing to the limit in a distinct way. An easy-to-implement bootstrap method is developed that allows for feasible implementation of the test. In an empirical application using high-frequency S&P 500 stock price data spanning the period from 2012 to 2021, we find evidence for 3 latent systematic jump factors that are not spanned by the traditional observable risk factors. (Joint work with Viktor Todorov, and Tony (Seunghyeon) Yu.)

Whitney Newey (Massachusetts Institute of Technology)

Title: Welfare Analysis in High Dimensional Dynamic Models

Abstract: This paper gives a consistent, asymptotically normal estimator of the expected value function when the state space is high-dimensional and the first-stage nuisance functions are estimated by modern machine learning tools. First, we show that value function is orthogonal to the conditional choice probability, therefore, this nuisance function needs to be estimated only at n−1/4 rate. Second, we give a correction term for the transition density of the state variable. The resulting orthogonal moment is robust to misspecification of the transition density and does not require this nuisance function to be consistently estimated. Third, we generalize this result by considering the weighted expected value. In this case, the orthogonal moment is doubly robust in the transition density and additional second-stage nuisance functions entering the correction term. We complete the asymptotic theory by providing bounds on second-order asymptotic terms. (Joint work with Victor Chernozhukov and Vira Semenova.)

Yacine Ait-Sahalia (Princeton University)

Title: So Many Jumps, So Few News

Abstract: This paper relates jumps in high frequency stock prices to firm-level, industry and macroeconomic news, in the form of machine-readable releases from Thomson Reuters News Analytics. We begin by examining the relationship from news to price jumps. We find that relevant news, both idiosyncratic and systematic, gets incorporated quickly into prices, as market efficiency suggests. However, in the reverse direction, the situation is different: the vast majority of price jumps do not have identifiable public news that can explain them. We then analyze the various market microstructure features that lead to jumps without news. (Joint work withChen Xu Li and Chenxu Li.)

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

Title: Nonparametric Standard Errors for High Frequency Data: The Continuous Time Observed Asymptotic Variance (C-AVAR)

Abstract: High frequency financial data has become an essential component of the digital world, giving rise to an increasing number of estimators. However, it is hard to reliably assess the uncertainty of such estimators. The Observed Asymptotic Variance (observed AVAR) is a non-parametric (squared) standard error for high-frequency-based estimators. We have earlier developed such an AVAR with time-discretization and two tuning parameters (per dimension). We here propose a better estimator C-AVAR by passing to continuous time. This is natural since observations are typically irregularly spaced. The C-AVAR only depends on a single tuning parameter delta, which permits a deeper analysis. We provide an expansion which shows that, when estimating the AVAR, there is a tradeoff (over a large range of delta between edge effect and the volatility of the spot parameter. The edge effect is given with exact order. The continuous-delta formulation is also useful in that it permits averaging across delta. This averaging is helpful (in data) because it smooths out edge effects. The averaging, as well as the continuous-time formulation, also means that the new estimator is more correct from a sufficiency standpoint. We show in a data illustration that the C-AVAR provides reasonable values for standard error when estimating integrated volatility in the presence of microstructure noise. As a by-product, the C-AVAR provides an interesting estimator of the volatility of volatility (or other spot process). As we shall also see in our

data illustration, the estimator is able to pick out important dates in recent financial history.

Jiashun Jin (Carnegie Mellon University)

Title: The Statistics Triangle

Abstract: In his Fisher's Lecture in 1996, Efron suggested that there is a philosophical triangle in statistics with "Bayesian", "Fisherian", and "Frequentist" being the three vertices, and many representative statistical methods can be viewed as a convex linear combination of the three philosophies. We collected and cleaned a data set consisting of the citation and bibtex (e.g., title, abstract, author information) data of 83,331 papers published in 36 journals in statistics and related fields, spanning 41 years. Using the data set, we constructed 21 co-citation networks, each for a time window between 1990 and 2015. We propose a dynamic Degree-Corrected Mixed-Membership (dynamic-DCMM) model, where we model the research interests of an author by a low-dimensional weight vector (called the network memberships) that evolves slowly over time. We propose dynamic-SCORE as a new approach to estimating the memberships. We discover a triangle in the spectral domain which we call the Statistical Triangle, and use it to visualize the research trajectories of individual authors. We interpret the three vertices of the triangle as the three primary research areas in statistics: "Bayes", "Biostatistics" and "Nonparametrics". The Statistical Triangle further splits into 15 sub- regions, which we interpret as the 15 representative sub-areas in statistics. These results provide useful insights over the research trend and behavior of statisticians.

Dacheng Xiu (University of Chicago)

Title: Can Machines Learn Weak Signals?

Abstract: In this study, we investigate the performance of several statistical and machine learning techniques including Ridge and Lasso in high-dimensional situations with extremely small signal-to-noise ratios. Our theoretical analysis revealed that Ridge regression effectively learns weak signals using an appropriate tuning parameter, while Lasso regression's performance fails to surpass that of a naive predictor benchmark, which relies on zero (historical mean). Furthermore, we demonstrate that cross-validation remains a reliable method for selecting tuning parameters in Ridge regression, and that its out-of-sample R-squared can be employed to measure the signal-to-noise ratio in the data generating process. Empirically, we analyze economic predictive regressions across six distinct datasets from macroeconomic, microeconomic, and financial domains, four of which exhibit weak signals where our asymptotic theory is particularly relevant.

Wei Biao Wu (University of Chicago)

Title: Asymptotics for Constant Step Size Stochastic Gradient Descent

Abstract: I will discuss a novel approach to understanding the behavior of Stochastic Gradient Descent (SGD) with constant step size by interpreting its evolution as a Markov chain. Unlike previous studies that rely on the Wasserstein distance, our approach leverages the functional dependence measure and explore the Geometric-Moment Contraction (GMC) property to capture the general asymptotic behavior of SGD in a more refined way. In particular, our approach allow SGD iterates to be non-stationary but asymptotically stationary over time, providing quenched versions of the central limit theorem and invariance principle valid for averaged SGD with any given starting point. These asymptotic results allow for the initialization of SGD with multiple distinct step

sizes, which is a widespread practice in the discipline. We subsequently show a Richardson-Romberg extrapolation with an improved bias representation to bring the estimates closer to the global optimum. We establish the existence of a stationary solution for the derivative SGD process under mild conditions, enhancing our understanding of the entire SGD procedure across varied step sizes. Lastly, we propose an efficient online method for estimating the long-run variance of SGD solutions. This aligns with the recursive nature of SGD, thereby facilitating fast and efficient computations. (Joint work with Jiaqi Li, Zhipeng Lou, and Stefan Richter.)

Simon Du (University of Washington, Seattle)

Title: How Over-Parameterization Slows Down Convergence of Gradient Descent

Abstract: We investigate how over-parameterization impacts the convergence behaviors of gradient descent through two examples. In the context of learning a single ReLU neuron, we prove that the convergence rate shifts from $exp(−T)$ in the exact-parameterization scenario to an exponentially slower $1/T^3$ rate in the over-parameterized setting. In the canonical matrix sensing problem, specifically for symmetric matrix sensing with symmetric parametrization, the convergence rate transitions from $exp(−T)$ in the exact-parameterization case to $1/T^2$ in the over-parameterized case. Interestingly, employing an asymmetric parameterization restores the $exp(−T)$ rate, though this rate also depends on the initialization scaling. Lastly, we demonstrate that incorporating an additional step within a single gradient descent iteration can achieve a convergence rate independent of the initialization scaling.

Cong Ma (University of Chicago)

Title: The Power of Preconditioning in Overparameterized Low-Rank Matrix Sensing

Abstract: We propose ScaledGD(λ), a preconditioned gradient descent method to tackle the low-rank matrix sensing problem when the true rank is unknown, and when the matrix is possibly ill-conditioned. Using overparametrized factor representations, ScaledGD(λ) starts from a small random initialization, and proceeds by gradient descent with a specific form of damped preconditioning to combat bad curvatures induced by overparameterization and ill-conditioning. At the expense of light computational overhead incurred by preconditioners, ScaledGD(λ) is remarkably robust to ill-conditioning compared to vanilla gradient descent (GD) even with overprameterization. Specifically, we show that, under the Gaussian design, ScaledGD(λ) converges to the true low-rank matrix at a constant linear rate after a small number of iterations that scales only logarithmically with respect to the condition number and the problem dimension. This significantly improves over the convergence rate of vanilla GD which suffers from a polynomial dependency on the condition number. Our work provides evidence on the power of preconditioning in accelerating the convergence without hurting generalization in overparameterized learning.

Jason Lee (Princeton University)

Title: Feature learning with gradient descent and smoothing

Abstract: We focus on the task of learning a single index model \sigma(w* x)$ with respect to the isotropic Gaussian distribution in d dimensions, including the special case when \sigma is a kth order hermite which corresponds to the Gaussian analog of parity learning. Prior work has shown that the sample complexity of

learning w* is governed by the \emph{information exponent} k* of the link function \sigma, which is defined as the index of the first nonzero Hermite coefficient of \sigma. Prior upper bounds have shown that n > d^{k*-1} samples suffice for learning w* and that this is tight for online SGD (Ben Arous et al., 2020). However, the CSQ lower bound for gradient based methods only shows that n > d^{k*/2}$ samples are necessary. In this work, we close the gap between the upper and lower bounds by showing that online SGD on a smoothed loss learns w* with n > d^{k*/2}$ samples. Next, we turn to the problem of learning multi index models f(x) = g(Ux), where U encodes a latent representation of low dimension. Significant prior work has established that neural networks trained by gradient descent behave like kernel methods, despite significantly worse empirical performance of kernel methods. However, in this work we demonstrate that for this large class of functions that there is a large gap between kernel methods and gradient descent on a two-layer neural network, by showing that gradient descent learns representations relevant to the target task. We also demonstrate that these representations allow for efficient transfer learning, which is impossible in the kernel regime. Specifically, we consider the problem of learning polynomials which depend on only a few relevant directions, i.e. of the form f*(x)=g(Ux) where U is d by r. When the degree of f* is p, it is known that n≍dp samples are necessary to learn f* in the kernel regime. Our primary result is that gradient descent learns a representation of the data which depends only on the directions relevant to f*. This results in an improved sample complexity of n≍d^2r+drp. Furthermore, in a transfer learning setup where the data distributions in the source and target domain share the same representation U but have different polynomial heads we show that a popular heuristic for transfer learning has a target sample complexity independent of d.

Yuqi Gu (Columbia University)

Title: Identifiable Deep Generative Models with Discrete Latent Layers

Abstract: We propose a class of identifiable deep generative models for very flexible data types. The key features of the proposed models include (a) discrete latent layers and (b) a shrinking pyramid- or ladder-shaped deep architecture. We establish model identifiability by developing transparent conditions on the sparsity structure of the deep generative graph. The proposed identifiability conditions can ensure estimation consistency in both the Bayesian and frequentist senses. As an illustration, we consider the two-latent-layer model and propose shrinkage estimation methods to recover the latent structure and model parameters. Simulation results corroborate the identifiability of the model, and also demonstrates the excellent empirical performance of our algorithm. Applications of the methodology to DNA nucleotide sequence data uncover useful discrete latent features that are highly predictive of held-out biological labels. The proposed framework provides a recipe for interpretable unsupervised learning and deep generative modeling.

Hongseok Namkoong (Columbia University)

Title: Adaptive Experimentation at Scale

Abstract: Standard bandit algorithms that assume continual reallocation of measurement effort are challenging to implement due to delayed feedback and infrastructural/organizational difficulties. Motivated by practical instances involving a handful of reallocation epochs in which outcomes are measured in batches, we develop a computation-driven adaptive experimentation framework that can flexibly handle batching. Our main observation is that normal approximations, which are universal in statistical inference, can also guide the design of adaptive algorithms. By deriving a Gaussian sequential experiment, we formulate a dynamic program that can leverage prior information on average rewards. Instead of the typical theory-driven paradigm, we leverage

computational tools and empirical benchmarking for algorithm development. Our approach significantly improves statistical power over standard methods, even when compared to Bayesian bandit algorithms (e.g., Thompson sampling) that require full distributional knowledge of individual rewards. Overall, we expand the scope of adaptive experimentation to settings that are difficult for standard methods, involving limited adaptivity, low signal-to-noise ratio, and unknown reward distributions.