Statistics in the Age of AI

May 9-11, 2024
The George Washington University

Sponsored By:

Organizing Committee

Xiaoke Zhang (Chair), Department of Statistics, George Washington University.

Kehui Chen, Department of Statistics, University of Pittsburgh.

Alexander Petersen, Department of Statistics, Brigham Young University.

Subrata Kundu, Department of Statistics, George Washington University.

Joshua Landon, Department of Statistics, George Washington University.

Gefei Lin, Department of Statistics, George Washington University.

Book of Talk Abstracts

Session I

Widespread Panic over Model Collapse

David Donoho
Stanford University

Modern ML systems are extraordinarily data hungry; and some major commercial players are said to now be using synthetic data to train their most ambitious ML systems. Also, AI-generated data will soon flood the internet, perhaps to the point where most available data are synthetic. Recently ML research has started to confront the larger issues that synthetic data might pose, including a future where most or all of the data available for an ML training are synthetic. A number of ML papers became prominent after promoting the idea of "model collapse", the "curse of recursion", "model autophagy disorder". Featuring experiments and some very basic theoretical argumentation they promoted a storyline where successive recycling of purely synthetic data led to model degeneration. Statisticians looking at the same setting as the ML researchers, have their own tradition and mental tools, and can provide a more balanced view of the situation. Depending on the synthetic data use case, no such collapse occurs. Empirical work with canonical LLMs and diffusion models confirms the absence of collapse, in the recommended use case.

Causal Learning from Heterogeneous Environments

Jianqing Fan
Princeton University

This talk develops causal learning from a multiple environments regression model in which data from heterogeneous experimental settings are collected. The joint distribution of the response variable and covariate may vary across different environments, yet the conditional expectation of outcome given the unknown set of important or quasi-causal variables is invariant across environments. We construct a novel environment invariant linear least squares (EILLS) objective function, a multiple-environment version of linear least squares that leverages the above conditional expectation invariance structure and heterogeneity among different environments to determine the true parameter. We establish non-asymptotic error bounds on the estimation error for the EILLS estimator in the presence of spurious variables. Moreover, we further show that the ℓ_0 penalized EILLS estimator can achieve variable selection consistency in high-dimensional regimes. These non-asymptotic results demonstrate the sample efficiency of the EILLS estimator and its capability to circumvent the curse of endogeneity algorithmically without any prior structural knowledge. The idea and concepts are further extended to the nonparametric model under heterogeneous environments. Leveraging the representation power of neural networks, we introduce neural causal networks based on a focus adversarial invariance regularization (FAIR) and its novel training algorithm. It is shown that the resulting procedure is adaptive to low-dimensional composition structures and can find causal variables under the structural causal models. The procedures are convincingly demonstrated using simulated examples.

Optimal Convex M-estimation via Score Matching

Richard Samworth
Cambridge University

In the context of linear regression, we construct a data-driven convex loss function with respect to which empirical risk minimization yields optimal asymptotic variance in the downstream estimation of the regression coefficients. Our semi-

parametric approach targets the best decreasing approximation of the derivative of the log-density of the noise distribution. At the population level, this fitting process is a nonparametric extension of score matching, corresponding to a log-concave projection of the noise distribution with respect to the Fisher divergence. The procedure is computationally efficient, and we prove guarantees on its asymptotic relative efficiency compared with an oracle procedure that has knowledge of the error distribution. As an example of a highly non-log-concave setting, for Cauchy errors, the optimal convex loss function is Huber-like, and yields an asymptotic relative efficiency greater than 0.87; in this sense, we obtain robustness without sacrificing (much) efficiency.

Session II

Object Oriented Data Analysis

Steven Marron
University of North Carolina, Chapel Hill

While Big Data is important, even greater challenges are being presented by Complex Data. OODA is a framework for addressing that. This talk considers the challenge of finding modes of variation (for data visualization) in the case of data that naturally lie on various curved manifolds.

Interpretable and Spatially-Aware Integration of Spatial Transcriptomics Datasets

Hongyu Zhao
Yale University

Recent advances in spatial transcriptomics technologies have led to growing diverse datasets, offering great opportunities to study tissue organizations and functions with spatial contexts. However, achieving precise and interpretable integration of these data originating from different samples, technologies, and developmental stages remains a challenge. In this presentation, I will present INSPIRE, a deep learning method for effectively integrating multiple spatial transcriptomics datasets. By incorporating non-negative matrix factorization, INSPIRE can uncover interpretable spatial factors with corresponding gene signatures, facilitating various downstream analyses. With designs of graph neural networks and an adversarial learning mechanism, INSPIRE enables spatially-aware and flexible integration of data from varying sources. We apply INSPIRE to analyze various datasets, including human cortex slices, mouse brain slices from different samples; mouse hippocampus slices, mouse embryo slices generated by four different technologies; and spatiotemporal atlases of mouse organogenesis containing half a million spatial spots. INSPIRE shows superior performance in capturing detailed biological signals, effectively borrowing information across different profiling technologies, and unraveling dynamical changes during embryonic development. In addition, we further apply INSPIRE to build 3D reconstruction of tissues using multiple slices, demonstrating its power and versatility.

Utilizing Synthetic Components to Balance Privacy Protection and Data Utility

Naisyin Wang
University of Michigan

The importance of privacy protection is rising in the current practice of publicly sharing data. Different evaluation criteria in terms of privacy protection and data utilities are considered. They may or may not agree with each other. In this presentation, we illustrate ways to utilize synthetic components to balance privacy protection and data utilities for different evaluation criteria. The main step is to enable a learning mechanism to preserve information in the privacy-oriented per-

turbed data in the generated observations for valid inferences. One aim is to enable users to easily implement statistical analysis using publicly shared data after the observations are processed with such privacy protection procedures. The efficacy and quality of the proposed procedures are illustrated theoretically and numerically via applications to biomedical datasets.

Session III

Autoregressive Networks with Dependent Edges

Qiwei Yao
London School of Economics and Political Science

We propose an autoregressive framework for modelling dynamic networks with dependent edges. It encompasses the models which accommodate, for example, transitivity, density-dependent and other stylized features often observed in real network data. By assuming the edges of network at each time are independent conditionally on their lagged values, the models, which exhibit a close connection with temporal ERGMs, facilitate both simulation and the maximum likelihood estimation in the straightforward manner. Due to the possible large number of parameters in the models, the initial MLEs may suffer from slow convergence rates. An improved estimator for each component parameter is proposed based on an iteration based on the projection which mitigates the impact of the other parameters. Based on a martingale difference structure, the asymptotic distribution of the improved estimator is derived without the stationarity assumption. The limiting distribution is not normal in general, and it reduces to normal when the underlying process satisfies some mixing conditions. Illustration with a transitivity model was carried out in both simulation and two real network data sets.

Boosting Data Analytics with Synthetic Volume Expansion

Xiaotong Shen
University of Minnesota

Synthetic data generation heralds a paradigm shift in data science, addressing the challenges of data scarcity and privacy and enabling unprecedented performance. As synthetic data gains prominence, questions arise regarding the accuracy of statistical methods compared to their application on raw data alone. Addressing this, we introduce the Synthetic Data Generation for Analytics framework, which applies statistical methods to high-fidelity synthetic data produced by advanced generative models like tabular diffusion models through knowledge transfer. These models, trained using raw data, are enriched with insights from relevant studies. A significant finding within this framework is the generational effect: the error of a statistical method initially decreases with the integration of synthetic data but may subsequently increase. This phenomenon, rooted in the complexities of replicating raw data distributions, introduces the "reflection point," an optimal threshold of synthetic data defined by specific error metrics. Through one data example, we demonstrate the effectiveness of this framework.

Label Correction of Crowdsourced Noisy Annotations with an Instance Dependent Noise Transition Model

Grace Yi
University of Western Ontario

The predictive ability of supervised learning algorithms hinges on the quality of annotated examples, whose labels often come from multiple crowdsourced annotators with diverse expertise. To aggregate noisy crowdsourced annotations, many existing methods employ an annotator-specific instance-independent noise transition matrix to characterize the labeling skills of each annotator. Learning an instance-dependent noise transition model, however, is challenging and remains relatively less explored. To address this problem, in this paper, we formulate the noise transition model in a Bayesian framework and subsequently design a new label correction algorithm. Specifically, we approximate the instance-dependent noise transition matrices using a Bayesian network with a hierarchical spike and slab prior. To theoretically characterize the distance between the noise transition model and the true instance-dependent noise transition matrix, we provide a posterior-concentration theorem that ensures the posterior consistency in terms of the Hellinger distance. We further formulate the label correction process as a hypothesis testing problem and propose a novel algorithm to infer the true label from the noisy annotations based on the pairwise likelihood ratio test. Moreover, we establish an information-theoretic bound on the Bayes error for the proposed method. We validate the effectiveness of our approach through experiments on benchmark and real-world datasets.

Session IV

Doubly Robust Uncertainty Quantification for Quantile Treatment Effects in Sequential Decision Making

Lan Wang
University of Miami

We consider multi-stage sequential decision-making, where the treatment at any stage may depend on the subject's treatment and covariate history up to that decision point. In this setting, we introduce a general framework for doubly robust uncertainty quantification for the quantiles of the cumulative outcome corresponding to a sequential treatment rule of interest, given the baseline covariate status. While previous literature has focused on the mean effects, quantile effects provide unique insights for understanding the distributional properties of the treatment effects and are more robust for heavy-tailed outcomes. Furthermore, unlike doubly robust estimation, doubly robust inference is substantially more challenging and remains largely unexplored for the quantile treatment effects, even in the single-stage setting. It is also known that for estimating the mean effects, a doubly robust estimation procedure does not automatically guarantee doubly robust inference. Our approach first yields a doubly robust estimator for an arbitrary quantile of interest based on pre-collected data, achieving semi-parametric efficiency. Next, we propose a novel doubly robust estimator for the asymptotic variance, facilitating the construction of a doubly robust confidence interval. To overcome the challenges associated with non-smoothness and parameter-dependent nuisance functions, we leverage empirical process techniques and deep conditional generative learning methods. We demonstrate the advantages of our approach via both simulation and a real data example from a short video platform. In addition, we observe that the proposed approach leads to an alternative mean effect estimator that outperforms existing estimators in dealing with heavy-tailed outcome distributions.

First-hitting-time Threshold Regression and Neural Network

Mei-Ling Ting Lee
University of Maryland, College Park

Disease progression in a patient can be described mathematically as a latent stochastic process. The patient experiences a failure event when his/her disease progression first reaches a critical threshold level. This happening defines a failure event as a first hitting time (FHT). First hitting time threshold regression (TR) models are based on an underlying stochastic process. The methodology does not require the proportional hazards assumption and represents a realistic alternative to the Cox model for capturing granular structure in a high-dimensional model. Machine learning methods such as boosting and neural networks have been applied to FHT TR models for prediction and causal inference. In addition to parametric models, FHT TR models have recently been extended to semiparametric applications.

Nonparametric Causal Additive Models with Smooth Backfitting

Byeong U. Park
Seoul National University

A fully nonparametric approach to learning causal structures from observational data is proposed. The method is described in the setting of additive structural equation models with a link to causal inference. The estimation procedure of the additive structural equation functions is based on a smooth backfitting approach. The flexibility of the nonparametric procedure results in strong theoretical properties in the estimation of the variable ordering. It is shown that under mild conditions the ordering estimate is consistent. Through simulations it is demonstrated that our method is superior to the state of the art approaches to causal learning. In particular, the smooth backfitting approach shows robustness when the noise is heteroscedastic.

Session V

Expectation Propagation and Maximum Likelihood in Generalized Linear Mixed Models

Iain Johnstone
Stanford University

We consider a class of generalized linear mixed models in which both the number of groups and the number of observations within each group are large, and in which usual likelihood analysis encounters both computational and technical challenges. Matt Wand and colleagues have adapted the machine learning technique of expectation propagation (EP) to yield state-of-the-art estimation of parameters in such models. Here we ask: are the EP estimators asymptotically efficient? A main challenge is to define an appropriate objective function that captures the EP iteration and approximates maximum likelihood well enough to inherit its efficiency. A second issue is to show that maximum likelihood actually is efficient, due to integrals over random effects in the likelihood. For this we propose a novel method based on the Vitali-Porter theorem of classical complex analysis.

Build an End-to-End Scalable Data Science Ecosystem Using Statistics, ML, and AI for Whole Genome Sequencing Analysis

Xihong Lin
Harvard University

Whole Genome/Exome Sequencing (WGS/WES) data and Electronic Health Records (EHRs), such as large scale national and institutional biobanks, have emerged rapidly worldwide. In this talk, I will provide an overview of building a data science end-to-end ecosystem of scalable analysis of large biobank- and population-based Whole Genome Sequencing (WGS) association studies of common and rare variants. I will discuss rare variant association tests (RVATs) using individual level data and RV meta-analysis methods using WGS summary statistics. These include STAAR and MetaSTAAR that improves the RVAT power by incorporating whole genome variant functional annotations and the ensemble (EN) tests. I will discuss improving scalability of WGS analysis by fitting mixed models using sparse GRM using individual-level data and sparse LD matrix using summary statistics. To build an end-to-end WGS data science ecosystem, I will introduce FAVOR (favor.genohub.org) and the LLM based FAVOR-GPT, a variant functional annotation online database and portal that provides multi-faceted functional annotations of genome-wide 9 billion variants, and FAVORAnnotator, a tool to functionally annotate any WGS/WES studies. Cloud-based platforms for these resources will be discussed. Results of large scale population-based multi-ancestry WGS studies and biobanks will be discussed, including the Trans-Omics Precision Medicine Program (TOPMed) from the National Heart, Lung and Blood Institute, the Genome Sequencing Program (GSP) of the National Human Genome Research Institute, the UK Biobank and All of Us. These studies have been collectively sequenced about 1.3 million genomes.

How Can Statistics Help Students Prepare for Better-Paid Jobs?

Runze Li
Penn State University

It is important to quantify the differences in returns to skills using the online job advertisements data, which have attracted great interest in both labor economics and statistics fields. This motivates us to study the relationship between the posted salary and the job requirements in online labor markets. In my talk, I will introduce new variable screening procedures for analysis of online job advertisements data and identify the important skill words related to the posted salary. I will illustrate the proposed procedures via an empirical study of the text data of job advertisements for data scientists and data analysts in a major China's online job posting website. This empirical analysis finds that the skill words like optimization, long short-term memory (LSTM), convolutional neural networks (CNN), collaborative filtering, are positively correlated with the salary while the words like Excel, Office, data collection, may negatively contribute to the salary.

Session VI

An RKHS Approach for Variable Selection in High-dimensional Functional Linear Models

Tailen Hsing
University of Michigan

High-dimensional functional data has become increasingly prevalent in modern applications such as high-frequency financial data and neuroimaging data analysis. We investigate a class of high-dimensional linear regression models, where each predictor is a random element in an infinite dimensional function space, and the number of functional predictors **p** can potentially be much greater than the sample size **n**. Assuming that each of the unknown coefficient functions belongs

to some Reproducing Kernel Hilbert Space (RKHS), we regularized the fitting of the model by imposing a group elastic-net type of penalty on the RKHS norms of the coefficient functions. We show that our loss function is Gateaux sub-differentiable, and our functional elastic-net estimator exists uniquely in the product RKHS. Under suitable sparsity assumptions and a functional version of the irrepresentible condition, we establish the variable selection consistency property of our approach.

Graphical Models in Infinite Dimensions

Victor Panaretos
École Polytechnique Fédérale de Lausanne

Graphical models allow us to distinguish direct and indirect associations in data. For the most part, one considers a collection of random vectors indexed by a finite or discrete set. The purpose of this talk is to explore what happens when we are interested in the associations within (intrinsic) and between (extrinsic) continuous time stochastic processes. In the first case, we are dealing with uncountably indexed real random variables. In the second case, we are dealing with discretely indexed Hilbertian random elements. Either setting introduces conceptual challenges owing to the lack of infinite-dimensional analogues of familiar algebraic tools, such as matrix inverses, factorisations, and determinants. We will see how the two problems share commonalities and notable differences, making contact with a classical problem in analysis: the continuation of positive-definite kernels.

Machine Learning with Functional Predictors and Applications to Crop Yield Prediction

Yehua Li
University of California, Riverside

Reliable prediction for crop yield is crucial for economic planning, food security monitoring, and agricultural risk management. This study aims to develop crop yield forecasting models at large spatial scales using trajectories of meteorological variables closely related to crop growth. The influence of climate patterns on agricultural productivity can be spatially inhomogeneous due to local soil and environmental conditions. We investigate several statistical machine learning methods (frequentist and Bayesian) to predict county-level corn yield for Midwestern states, using multivariate functional predictors such as precipitation and temperature trajectories. Our study provides further insights into understanding the impact of climate change on crop yield.

Session VII

Integrative Deep Multi Learning for Biclustering and Predicting Cancer Drug Responses: Leveraging Omics and Drug Molecular Data

Haiyan Huang
University of California, Berkeley

Precision medicine in cancer treatment leverages the complex relationship between cancer biology and drug molecules, hindered by the genetic complexity of cancer and diverse drug structures. We introduce Integrative Multi Task Deep Biclustering (IMTDB), merging cancer omics, drug data, and drug response to pinpoint cancer cell lines sensitive to specific drugs based on their molecular profiles. IMTDB uses biclustering to identify these sensitive subsets and predict

drug sensitivity with enhanced accuracy by iterating between learning cell line and drug embeddings and their response mappings. This approach helps identify tailored treatment strategies by revealing the molecular signatures driving drug response. IMTDB's capability to group unseen cell lines and compounds facilitates quick screening, marking a potential significant step towards personalized cancer therapy. Our validation through simulations and diverse datasets underscores IMTDB's potential in identifying precise treatment options.

Functional Regression through Distributed Learning:
An Application to Brain Imaging Studies

Lily Wang
George Mason University

Motivated by recent work analyzing data in biomedical imaging studies, we consider a class of functional regression models for imaging responses. We introduce a novel nonparametric distributed (NPD) learning framework that utilizes multivariate spline smoothing over a triangulation of domain. The proposed NPD estimation algorithm features a scalable and communication-efficient implementation scheme to achieve near-linear speedup. Asymptotic confidence intervals and data-driven simultaneous confidence corridors (SCCs) for the coefficient functions are constructed. Our method can simultaneously estimate and make inferences of the coefficient functions while incorporating the spatial heterogeneity. In addition, we provide rigorous theoretical support for the NPD estimation and inference framework. Specifically, we demonstrate that the NPD-based spline estimators are asymptotic normal, and have the same convergence rate as the global spline estimators obtained using the entire dataset. Monte Carlo simulation studies are conducted to examine the finite-sample performance of the proposed method. The proposed method is applied to the spatially normalized Positron Emission Tomography (PET) data of Alzheimer's Disease Neuroimaging Initiative (ADNI).

A Supervised Deep Learning Method for Nonparametric Density Estimation

Johannes Schmidt-Hieber
University of Twente

Nonparametric density estimation is an unsupervised learning problem. In this work we propose a two-step procedure that casts the density estimation problem in the first step into a supervised regression problem. The advantage is that we can afterwards apply supervised learning methods. Compared to the standard nonparametric regression setting, the proposed procedure creates, however, dependence among the training samples. To derive statistical risk bounds, one can therefore not rely on the well-developed theory for i.i.d. data. To overcome this, we prove an oracle inequality for this specific form of data dependence. As an application, it is shown that under a compositional structure assumption on the underlying density, the proposed two-step method achieves convergence rates that are faster than the standard nonparametric rates. A simulation study illustrates the finite sample performance.