The 39th New England Statistics Symposium

From Data to Discovery: Statistical Insight in the Age of AI

Conference Dates and Location: May 27–29, 2026, Storrs, CT

Program Book

Contents

Parallel Session 1 | 08:30 AM - 10:10 AM, May 28

Advances in Mathematical Finance (IS-4)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 201

Chair: Oleksii Mostovyi

Proposer: Oleksii Mostovyi, University of Connecticut

Presenters: Gu Wang, Oleksii Mostovyi, Xiaohang Ma, Yaacov Kopeliovich

Continuous Policy and Value Iteration for Stochastic Control Problems and Its Convergence

Gu Wang, Worcester Polytechnic Institute

Abstract. We introduce a continuous policy-value iteration algorithm where the approximations of the value function of a stochastic control problem and the optimal control are simultaneously updated through Langevin-type dynamics. This framework applies to both the entropy-regularized relaxed control problems and the classical control problems, with infinite horizon. We establish policy improvement and demonstrate convergence to the optimal control under the monotonicity condition of the Hamiltonian. By utilizing Langevin-type stochastic differential equations for continuous updates along the policy iteration direction, our approach enables the use of distribution sampling and non-convex learning techniques in machine learning to optimize the value function and identify the optimal control simultaneously.

The Value of Partial Information

Oleksii Mostovyi, University of Connecticut

Abstract. We investigate a pricing rule that is applicable for streams of income or contingent claim liabilities and study how this rule changes under additional insider-type information that an investor might obtain. Considering a model where the risky asset might have jumps, we obtain an explicit form of the associated state price density for the three different types of agents: one who has no information about the jumps, one who knows in advance exactly when the each jump will occur, and one who has no information about the size of the jumps but has partial information about the size of each jump. For each of these agents, we provide characterizations of the pricing rule and establish a representation formula, allowing us to quantify the value of partial information for streams of labor income or contingent claim liabilities. Our work is motivated by finding and characterizing a pricing rule that, both with or without partial information about jumps, assigns different values of information for different income streams or contingent claim liabilities. Based on the joint work with Philip Ernst (Imperial College London).

Duality for optimal stopping in continuous time

Xiaohang Ma, University of Connecticut

Abstract. We character the duality of optimal stopping problem in the Lagrange-Mayer formulation. The key contribution is the construction of a dual problem of a mixed optimal stopping and singular stochastic control type, which allows us to bypass the inherent lack of convexity in the set of stopping times. Specifically, we establish the dual characterization of the value function, provide a dual representation of the Snell envelope in the Mayer case, and introduce counterexamples that motivate the necessary technical assumptions.

Merton portfolio problem under Isoelastic Utility with exogenous bankruptcy

Yaacov Kopeliovich, University of Connecticut

Abstract. Solution for Merton portfolio problem for power and log utilities is presented under exogenous bankruptcy for stock. As a result we obtain a simple example of non-myopic weights for optimal allocation of assets.

Use of Bayesian Statistics in Clinical Development (IS-30)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 202

Chair: Vickie (Yuanye) Zhang

Proposer: Vickie (Yuanye) Zhang, Servier Pharmaceuticals

Presenters: Zhaohua Lu, Ming-Dauh Wang, Alex Sverdlov

Bayesian Hybrid Design and Practical Considerations

Zhaohua Lu, Daiichi-Sankyo Inc.

Abstract. Early-phase drug development for combination therapies aims to preliminarily assess additive activity when a novel agent is combined with an established monotherapy. While uncontrolled single-arm trials are commonly used due to feasibility constraints, these designs often face challenges in advancing to subsequent phases because of the absence of randomization and limited ability to robustly evaluate efficacy. Hybrid trial designs, which borrow external information from completed historical studies, offer a promising alternative by enhancing statistical power and study efficiency. The Bayesian dynamic power prior (DPP) framework enables flexible information borrowing, allowing customization to specific study needs and providing computational advantages through closed-form posterior distributions. The DPP approach, if only relies on marginal outcome similarity, can be problematic when baseline prognostic covariates differ between historical and concurrent populations, leading to residual confounding and inflated Type I error rates. To address this, a propensity score integrated DPP (PS-DPP) framework is proposed. PS-DPP employs a two-stage process: first, historical controls are reweighted or matched to the concurrent population using propensity scores; second, a global, gated, and dynamic borrowing architecture is applied, ensuring information borrowing is conditional on both covariate balance and outcome similarity. This dual-layer approach mitigates prior-data conflict and improves error control, particularly under population drift. Simulation studies and case analyses demonstrate that PS-DPP delivers superior operating characteristics compared to standard DPP and other propensity score methods, while maintaining computational efficiency and regulatory transparency. PS-DPP provides a robust, interpretable framework for Bayesian evidence synthesis in early-phase clinical trials, facilitating efficient study designs and more reliable decision-making when leveraging historical or real-world data.

Bayesian Quantitative Decision Making (QDM) for Late-Phase Event-Driven Clinical Trials

Ming-Dauh Wang, Bayer

Abstract. Late-phase event-driven trials are conventionally designed assuming a fixed treatment effect (e.g. hazard ratio = 0.8), often without sufficient evidence of the assumption. But, in reality our knowledge of the expected treatment effect is uncertain and can be informed by existing data. The long adopted deficient practice assuming fixed treatment effect, partly attributable to the established approach to sample size determination in the protocol for event-driven trials, could result in inadequately powered trials and lower than expected probabilities of trial success. We propose a novel evidence-driven QDM method for the design of late-phase event-driven trials to enhance trial success prediction. Our approach first summarizes early-phase short-treatment effect, often on a biomarker, of the investigational treatment by Bayesian analysis. Next, the summarized early-phase treatment effect (including uncertainty) is mapped to a late-phase clinical effect on the reduction of the hazard of the clinical event of interest by the treatment. The effect mapping is based on Bayesian elicitation of the correlation of the early-phase and late-phase treatment effects using internal and/or external information. Finally, the resulting late-phase clinical effect on hazard reduction, expressed as a predictive distribution, is used to calibrate the probability of trial success and to adjust the design assumptions toward meeting the desired probability. We will present applications of the proposed QDM approach to clinical programs, where investigational treatments reached the end-of-phase-1 or end-of-phase 2 decision point of entering the late-phase development. QDM analysis enabled quality milestone decisions and increased the reliability of the predicted levels of success for the late-phase trials aiming for regulatory submissions.

Recent statistical developments in modern biomedical and computational applications (IS-31)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 205

Chair: Pei Geng

Proposer: Pei Geng, University of New Hampshire

Presenters: Pei Geng, Chameli Piyatilake, Qing Wang, Liangliang Zhang

A two-stage GAN-based instrumental variable method for causal analysis of omics data

Pei Geng, University of New Hampshire

Abstract. Mendelian randomization (MR) utilizes

genetic variants as instrumental variables (IVs) to estimate the causal effects of disease-associated genes, thereby establishing putative causal associations and reducing spurious association findings due to confounding. To mitigate the potential bias due to the violation of IV conditions and nonlinear exposure–outcome relations in MR studies, we propose a two-stage deep learning framework, which is free from distribution assumptions of exposure given IVs and flexible to capture complex exposure–outcome relations. Specifically, we adapt the generative adversarial networks (GAN) to estimate the conditional distribution of gene expression given IVs in the first stage and apply deep functional neural networks to learn the causal relationships between gene expression and outcomes. Through simulation studies under different distributions and model choices, our proposed GAN-based instrumental variable (GAN-IV) method demonstrates improved performance over the two-stage least squares method, pleiotropy-robust MR methods (e.g.MR-LINK), and state-of-the-art deep-learning-based methods (e.g.DeLIVR). A real data application on the ROSMAP dataset further illustrates that GAN-IV is capable of capturing the exposure distribution and complex nonlinear causal effect between gene expression and disease phenotype.

Variable Selection in Case-Control Logistic Regression via Density Estimation

Chameli Piyatilake, University of New Hampshire

Abstract. Logistic regression is a fundamental tool for modeling binary outcomes in case-control studies, with broad applications in epidemiology and biomedical research. However, when the number of predictors is large or the covariates are correlated, classical penalized logistic regression methods can struggle with variable-selection stability. In this work, we build on the Integrated Squared Distance (ISD) framework of Geng and Sakhanenko (2016) to develop a new variable selection procedure for case-control logistic regression. Our approach embeds an elastic net penalty directly into the ISD estimation criterion, yielding an ISD-elastic net method that simultaneously estimates regression coefficients and selects relevant predictors. We assess the proposed method through simulation studies covering a range of realistic scenarios, including balanced and unbalanced case-control designs, independent and correlated covariate structures, and Gaussian, exponential, and gamma covariate distributions. Across these settings, the ISD-elastic net demonstrates stronger variable selection performance than classical penalized logistic regression, with gains most pronounced when covariates are highly correlated. Ongoing work extends the ISD-based penalized framework to measurement-error settings, where some or all predictors are contaminated. This extension aims to develop robust variable selection procedures for case-control studies in which covariate measurement error can otherwise lead to biased estimates and poor variable-selection performance.

Enhancing diversity and improving prediction performance of subsampling-based ensemble methods

Qing (Wendy) Wang, Wellesley College

Abstract. This talk discusses how diversity among training samples impacts the predictive performance of a subsampling-based ensemble. It is well known that diverse training samples improve ensemble predictions, and smaller subsampling rates naturally lead to enhanced diversity. However, this approach of achieving a higher degree of diversity often comes with the cost of a reduced training sample size, which is undesirable. This paper introduces two novel subsampling strategies–partition and shift subsampling–as alternative schemes designed to improve diversity without sacrificing the training sample size in subsampling-based ensemble methods. From a probabilistic perspective, we investigate their impact on subsample diversity when utilized with tree-based sub-ensemble learners in comparison to the benchmark random subsampling. Through extensive simulations and real-world examples in both regression and classification contexts, we found a significant improvement in the predictive performance of the developed methods. Notably, this gain is particularly pronounced on challenging datasets or when higher subsampling rates are employed.

McDisCov: outcome-linked MiCrobiome Discovery via Compositional lOgit-normal modeling under Varying support

Liangliang Zhang, Case Western Reserve University

Abstract. Microbial association studies are essential for understanding how the microbiome contributes to health and disease. Yet next-generation sequencing produces read counts constrained by sequencing depth rather than true microbial abundances, leading researchers to rely on two main strategies: composition data approaches for relative abundance

(RA) and count-based methods for absolute abundance (AA). Both frameworks require normalization and have inherent limitations, and analyses of the same data often yield divergent results, underscoring a lack of consensus and reproducibility. To address this gap, we propose a novel frame- work that unifies both RA-based and AA-based inferences in a compositional logit-normal model. Unlike existing methods, it eliminates the need for size factor estimation and zero imputation, both of which can introduce bias, through newly developed statistical techniques. This enables flexible RA inference with respect to arbitrary weighted reference groups and simultaneously allow AA inference through pairwise comparisons. Extensive simulations across diverse scenarios show that our method achieves consistently higher power and lower false dis- covery rates compared to widely used approaches such as ANCOM-BC2, LinDA and DESeq2. Applications to multiple real-world datasets further demonstrate its value in yielding reproducible biological discoveries with direct implications for advancing microbiome-based diagnostics and therapeutics.

Recent Developments in Regression, Causal Learning, and Multiple Testing (IS-33)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 206

Chair: Yang Ning

Proposer: Yang Ning, Cornell University

Presenters: Jian Yan, Yihong Gu, Cheng Yong Tang, Yang Ning

Machine-Learning-Assisted Comparison of Regression Functions

Jian Yan, Cornell University

Abstract. We revisit the classical problem of comparing regression functions, a fundamental question in statistical inference with broad relevance to modern applications such as data integration, transfer learning, and causal inference. Existing approaches typically rely on smoothing techniques and are thus hindered by the curse of dimensionality. We propose a generalized notion of kernel-based conditional mean dependence that provides a new characterization of the null hypothesis of equal regression functions. Building on this reformulation, we develop two novel tests that leverage modern machine learning methods for flexible estimation. We establish the asymptotic properties of the test statistics, which hold under both fixed- and high-dimensional regimes. Unlike existing methods that often require restrictive distributional assumptions, our framework only imposes mild moment conditions.

Is double machine learning optimal for black-box models?

Yihong Gu, Harvard University

Abstract. Modern semiparametric estimation often relies on flexible "black-box" machine learning methods to estimate nuisance functions, raising a fundamental question: how do nuisance estimation errors propagate into inference for low-dimensional target parameters? The dominant paradigm, exemplified by double/debiased machine learning (DML), yields error bounds in which nuisance estimation errors enter multiplicatively. While widely adopted, it remains unclear whether this product-rate dependence is optimal for black-box models. In this paper, we start by revisiting the partial linear model$Y = \mu_0(X) + T \cdot \beta_0 + \varepsilon$under a structure-agnostic setting, where the nuisance function$\mu_0$is estimated using a generic machine learning model, with approximation error$\delta_\mu^{\text{appr}}$and stochastic errorδ_μ^{stoc}. We show that the standard DML rate is not optimal in the regime where the auxiliary function$\mathbb{E}[T|X = x]$cannot be consistently estimated. We propose a new estimator for$\beta_0$that achieves a sharper rate of$n^{-1/2} + \delta_\mu^{\text{appr}} + (\delta_\mu^{\text{stoc}})^2$and establish a matching lower bound demonstrating its optimality. Our results reveal a new principle: the first-order stochastic error of nuisance estimation can be eliminated without imposing any additional assumptions. This also leads to a revised tuning strategy favoring under-smoothing, where$\delta_\mu^{\text{appr}} \asymp (\delta_\mu^{\text{stoc}})^2$, rather than the classical bias-variance trade-off with$\delta_\mu^{\text{appr}} \asymp \delta_\mu^{\text{stoc}}$. Under mild additional conditions, the estimator is asymptotically normal and semi-parametrically efficient. The result applies to neural network models to yield a practical solution to the efficient semiparametric estimation with multi- and high-dimensional nonparametric nuisance functions. The proposed method extends to a broad class of semi-parametric linear functional estimation problems, including average treatment effect estimation. Our results imply that popular orthogonal score methods in semiparametric estimation with black-box nuisance learners can be substantially improved.

Controlling the False Discovery Rate in High-Dimensional Linear Models Using Model-X Knockoffs and p-values

Cheng Yong Tang, Temple University

Abstract. We propose a novel multiple testing methodology for controlling the false discovery rate (FDR) in high-dimensional linear models that integrates model-X knockoff techniques with debiased penalized regression estimators. At the foundation of our methodology, we construct and study two sets of naturally paired high-dimensional test statistics and the associated p-values for evaluating the same null hypotheses. The first set is shown to be asymptotically mutually independent, justifying the use of the Benjamini-Hochberg procedure. We further exploit the pairing structure through a two-step procedure aimed at improving power. Our theoretical results establish the key properties of the framework with respect to asymptotic FDR control and formally characterize the associated power gains of the two-step procedure. Importantly, our framework accommodates general dependence in the design matrix. Extensive simulations demonstrate that our methods outperform existing approaches—particularly those relying on empirical FDP estimates—in both power and FDR control accuracy, with notable gains in settings involving weaker signals, small sample sizes, or low target FDR levels.

Active Subsampling for Binary Response Models

Yang Ning, Cornell University

Abstract. In the measurement-constrained problems, despite the availability of large datasets, we may be only affordable to observe the labels on a small portion of the large dataset. This poses a critical question that which data points are most beneficial to label given a budget constraint. In this paper, we focus on the estimation of the optimal individualized threshold in a measurement-constrained M-estimation framework. Our goal is to estimate a high-dimensional parameter in a linear threshold for a continuous variable X such that the discrepancy between whether X exceeds the threshold and a binary outcome is minimized. We propose a novel K-step active subsampling algorithm to estimate , which iteratively samples the most informative observations and solves a regularized M-estimator. We show that the two-step algorithm yields an estimator with the parametric convergence rate. Furthermore, we formulate a minimax framework for the measurement-constrained M-estimation problem and prove that our estimator is minimax rate optimal up to a logarithmic factor. Finally, we demonstrate the performance of our method in simulation studies and apply the method to analyze a large diabetes dataset.

Modeling Dependence, Dynamics, and Structure in Complex Data (IS-37)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 301

Chair: Shariq Mohammed

Proposer: Shariq Mohammed, Boston University

Presenters: Fatima Tuz-Zahra, Damla Ilter-Fakhouri, Adlin Pinheiro, Neil Spencer

Digital Biomarkers from Pen Trajectory Deviations in the Digital Trail Making Test for Detecting Mild Cognitive Impairment and Dementia

Fatima Tuz-Zahra, Boston University

Abstract. Digital cognitive assessments, such as the digital Trail Making Test (dTMT), generate densely sampled pen trajectory data that may contain richer information about cognitive status than traditional paper-based summary measures alone (Dahmen et al., 2017; Zhang et al., 2023). We investigated whether time-series features extracted from the TMT-A path deviation relative to the shortest path improve prediction of mild cognitive impairment (MCI) and dementia beyond conventional measures. Data were drawn from participants aged 60 and older (N=925; 43 cognitively impaired or demented) in the Framingham Heart Study. Pen trajectory data from the dTMT Part A were recorded at approximately 13-millisecond intervals. For each participant, we computed deviations from the shortest path between consecutive numbered circles, producing a densely sampled time series of deviations over the course of the test. We applied the Highly Comparative Time-Series Analysis (hctsa) feature set (Fulcher & Jones, 2017; Fulcher et al., 2013), extracting 7,755 time-series features capturing autocorrelation structure, fractal scaling, visibility graph properties, surrogate-based nonlinear statistics, preprocessing comparisons, forecasting characteristics, and others. After removing failed or zero-variance features (N=2,139 removed), we evaluated 5,616 candidate features using logistic regression with 5-fold cross-validation.

Our baseline model included age, sex, education, APOE 4 status, and total test completion time. Each candidate feature was individually added to this baseline model. Features with Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) less than that of the baseline model (AUC=0.82) plus a 2% relevance buffer were discarded. To eliminate redundancy among the 121 remaining features, k-medoid clustering identified 16 representative features from distinct correlation groups. The aim was to identify features that complement the baseline model and reveal distinct behavioral patterns in test performance. Stability selection with 1,000 bootstrap samples (Meinshausen & Bühlmann, 2010) identified five stable predictors. Finally, to assess variable importance and direction of effect, we fit a logistic regression model including all five features on the full cohort. The five stable predictors were: time-series surrogates (SD_TSTL_surrogates, OR=0.49, 95% CI: 0.32–0.73), spline-based preprocessing comparisons measuring stationarity (PP_Compare–stationarity, OR=0.52, 95% CI: 0.36–0.75), visibility graph power-law fit (NW_VisibilityGraph, OR=1.88, 95% CI: 1.14–3.10), autocorrelation shape (CO_AutoCorrShape, OR=0.61, 95% CI: 0.40–0.93), and spline-based preprocessing comparisons assessing Gaussianity of the deviation distribution (PP_Compare–distribution, OR=1.59, 95% CI: 1.15–2.20). These findings demonstrate that time-series features from digital pen trajectories capture aspects of cognitive performance not reflected in total completion time. Novel patterns were revealed, including: time-reversal asymmetry, reflecting how quickly errors accumulate relative to how quickly they are corrected (SD_TSTL_surrogates); variation in mean performance level across the test (PP_Compare–stationarity); whether the range of error magnitudes follows a broad, heavy-tailed distribution or is punctuated by occasional spikes (NW_VisibilityGraph); error persistence, capturing how long deviations remain elevated before returning to baseline (CO_AutoCorrShape); and whether the test exhibits a slow trend suggesting learning or fatigue versus being driven by error spikes (PP_Compare–distribution). These novel patterns offer potential digital biomarkers for early detection of cognitive decline, moving beyond traditional paper-based summary measures to characterize the temporal dynamics of cognitive performance.

Adaptive Statistical Strategies: Integrating Multi-Scale Learning Across Heterogeneous and Complex Data Structures

Damla ILTER FAKHOURI, CURRY COLLEGE / DEPARTMENT OF SCIENCE and MATHEMATICS

Abstract. In the era of high-dimensional data, the ability to synthesize information across varying scales and formats is a cornerstone of impactful statistical research. This presentation explores the development and implementation of adaptive statistical strategies designed to navigate the challenges inherent in heterogeneous and complex data structures. The session will delve into the methodological nuances of handling high-dimensional feature spaces, spatial dependencies, and longitudinal complexities. By utilizing adaptive regularization and hybrid machine learning techniques, we demonstrate how to maintain model robustness despite sparse sampling and data noise. Fur-thermore, the talk will transition from theoretical formulations to practical insights, drawing on experiences from two distinct domains: large-scale ecological monitoring and specialized oncological research. To conclude, we will critically evaluate the trade-offs between model complexity and interpretability. We will discuss the practical difficulties of data integration and the strategic advantages of cross disciplinary method-ological transfers, illustrating how adaptive frameworks can turn data heterogeneity from a barrier into a source of predictive strength.

Lifetime vascular risk factor trajectories and late-life cerebral small vessel disease burden

Adlin Pinheiro, Boston University School of Public Health

Abstract. Introduction: Cerebral small vessel disease (CSVD) is strongly associated with stroke and dementia risk, often developing years before onset of symptoms. Preventive efforts focus on treatment of modifiable vascular risk factors (VRF), but the complex relationship between changes in VRF over time and CSVD burden remain unclear. Thus, we studied how lifetime trajectories of VRFs relate to CSVD burden later in life. Methods: We included 7961 Framingham Heart Study participants with six or more repeated VRF assessments over their lifetime. Functional principal components (FPC) analysis was used to characterize VRF trajectories across the sample. Among those who underwent MRI, a multi-marker score quantifying CSVD burden assigned one point to each of the following:

cerebral microbleeds, covert infarcts, extensive white matter hyperintensities, cortical superficial siderosis and high burden of perivascular spaces, categorized as 0, 1, 2+ for analysis. Ordinal logistic regression was used to relate FPC scores (informed by the entire sample) of each VRF to CSVD score. Additionally, k-means clustering was used on the FPC scores to identify three to four distinct lifetime VRF trajectory groups, which were then examined in relation to CSVD burden. Results: In 1625 participants with CSVD measurements (mean age 73 ± 9 years, 45% male), 47% had a CSVD score of 0, 31% a score of 1, and 22% a score of 2+. Lifetime trajectories of systolic and diastolic blood pressure, pulse pressure, and cigarettes per day were significantly associated with greater CSVD burden. Cluster assignments for trends of systolic blood pressure, pulse pressure, and BMI were also significantly associated with greater CSVD burden. Conclusion: Lifetime patterns of blood pressure and smoking are associated with CSVD burden later in life. These findings suggest that preventive efforts should consider VRF trajectories, rather than single mid-life or late-life measurements, to target high risk individuals.

Bayesian Ordinal Probit Modeling of Multi-view Directed Social Networks

Neil A. Spencer, University of Connecticut

Abstract. Social network data are often collected through multiple relational "views," such as friendship, resource sharing, kinship, and reputation ratings (e.g., how generous is person A according to person B?). A central goal in analyzing such data is understanding how the views relate—for example, are individuals more likely to share resources with those they rate as generous? I develop a Bayesian modeling framework for such data, motivated by a large anthropological study of a rural Nicaraguan community. The dyadic ties collected in this setting are both directed and ordinal. To address the associated methodological and computational challenges, I adapt a multivariate ordinal probit model to the network setting and derive a tractable Bayesian inference strategy. Exchangeability of dyad members imposes nontrivial structural constraints on the latent correlation matrix; I introduce a new parametrization that satisfies these constraints, supports flexible prior specification, and enables efficient Bayesian inference. The result is a practical modeling approach that is straightforward to implement using popular Bayesian computation tools such as Stan.

Novel statistical modeling and computing methods for complex data (IS-38)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 305

Chair: victor hugo lachos davila

Proposer: victor lachos davila, University of Connecticut

Presenters: Mary Lai Salvana, Fusheng Yang, Aditya Vikram Sett, Dashun Liu

Power-Law Regression with Mixed Effects for Self-Organized Criticality

Mary Salvana, University of Connecticut

Abstract. Extreme value theory (EVT) models rare events using only observations above a high threshold, discarding most of the data. For processes exhibiting self-organized criticality (SOC)—such as power outages, precipitation, earthquakes, wildfires, and storm sizes—this is unnecessary: power laws govern the full distribution, not just the tail. Thus, observations far below EVT thresholds still inform tail behavior, with the power-law exponentα > 0determining the frequency of extremes across the entire support. Despite this, methods for incorporating covariate effects into power-law behavior remain underdeveloped, particularly frameworks that allow$\alpha$to vary with predictors. We address this gap by developing a mixed-effects regression model in whichαdepends on covariates, enabling flexible, context-sensitive inference on how external factors shape the frequency and severity of extremes. Simulations and applications show that the proposed model yields sharper tail inference for SOC processes than models without covariates and conventional EVT approaches.

Bayesian Comparison of Negative Binomial and Generalized Poisson INAR(1) Models for Zero-Inflated and Overdispersed Time Series

Fusheng Yang, University of Connecticut

Abstract. Modeling count time series is a critical task in many fields, particularly when data exhibit common characteristics such as overdispersion and an excess of zeros. While the first-order integer-valued autoregressive (INAR(1)) model provides a fundamental framework, standard versions often fail to capture these complex features. This study provides

a comprehensive comparison of two advanced extensions based on zero-inflated Negative Binomial and zero-inflated Generalized Poisson innovations. We employ a unified Bayesian framework using Hamiltonian Monte Carlo in Stan for parameter estimation and model comparison. The performance of these two approaches is rigorously assessed through a detailed simulation study that includes scenarios in which each is the true data-generating process, as well as a model-misspecification scenario to assess robustness. Our simulation results indicate that, while both specifications perform well when the model is correctly identified, the model based on the Generalized Poisson distribution exhibits greater robustness and flexibility when the underlying data distribution is unknown. An application to two real-world datasets, monthly drug offenses and monthly sex offenses, further supports these findings. Despite the differing underlying characteristics of the two series, model selection criteria consistently favor the Generalized Poisson variant. To support reproducibility and broader adoption, we provide the updated open-source R package, ZIHINAR1, which implements this unified Bayesian modeling framework. These results offer practical guidance for researchers and practitioners in selecting appropriate models for complex, overdispersed, and zero-inflated count time series.

Bayesian mixed-effects censored regression based on the odd log-logistic Weibull distribution

Aditya Vikram Sett, University of Connecticut

Abstract. In the agricultural and biological sciences, evaluating the time required for livestock to reach critical developmental milestones, such as an optimal slaughter weight, frequently involves clustered, right-censored data with complex risk profiles. Standard mixed-effects survival models typically assume constant or strictly monotonic hazard rates, which are fundamentally inadequate for capturing the unimodal, "hump-shaped'' growth dynamics characteristic of physiological development. To address this methodological gap, we propose a highly flexible Bayesian mixed-effects censored regression model based on the odd log-logistic Weibull (OLLW) distribution. The proposed framework naturally accommodates non-monotonic hazard rates while accounting for residual genetic correlation via multivariate normal random effects. Parameter estimation is performed using Markov chain Monte Carlo (MCMC) methods. Comprehensive simulation studies confirm the robust recovery of the model parameters across varying group structures, sample sizes, and censoring proportions. Finally, the applicability of the proposed model is illustrated through the analysis of a real dataset from the animal genetic improvement sector, evaluating the time required for Nelore calves to reach a 160 kg target weight. Results demonstrate that the proposed framework provides a superior fit compared to standard distributions, providing practitioners with a probabilistic, biologically interpretable tool for ranking sire genetic superiority.

A unified EM framework for high-dimensional linear mixed-effects models with regularization penalties

dashun liu, University of Connecticut

Abstract. Linear mixed-effects models (LMMs) provide a fundamental framework for analyzing longitudinal and clustered data across various scientific disciplines. However, the increasing prevalence of modern data collection methods frequently yields datasets where the number of predictors is comparable to much larger than the number of subjects ($p \gg n$). In such high-dimensional settings, variable selection is critical, yet standard likelihood-based procedures become highly unstable and computationally prohibitive due to the complex parameter space and the presence of unobserved random effects. To overcome these challenges, we propose a flexible Expectation-Maximization (EM) based framework for penalized maximum likelihood estimation that seamlessly accommodates a broad class of regularization penalties. By treating random effects as missing data, the E-step computes their conditional moments to construct a pseudo-response. The M-step reduces the optimization to a sequence of penalized least-squares problems, which can be solved efficiently using fast coordinate-descent algorithms available in R packages such as `glmnet` and `ncvreg`. Our unified framework encompasses the Least Absolute Shrinkage and Selection Operator (Lasso), Adaptive Lasso (ALasso), and nonconvex penalties, including Smoothly Clipped Absolute Deviation (SCAD) and Minimax Concave Penalty (MCP). Tuning parameters are selected via cross-validation embedded within the EM iterations. Simulation studies across various dimensional regimes and correlation structures demonstrate the algorithm's effectiveness in both variable selection accuracy and parameter estimation. Finally, we apply the proposed methodology to a riboflavin gene expression dataset, in which nonconvex penalties successfully identify a

sparse set of biologically plausible genes associated with riboflavin production.

Novel Statistical Methods for Diverse Applications (IS-51)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 101

Chair: Maryclare Griffin

Proposer: Maryclare Griffin, University of Massachusetts Amherst

Presenters: David Burt, Rebecca Kurtz-Garcia, Troy Wixson, Abbas Zaidi

Bayesian Modeling for Battle Royale Style Competitions

Rebecca Kurtz-Garcia, Smith College

Abstract. In recent years there has been a large rise of battle royale tournaments. In this style tournament competitors are pitted against each and eliminated sequentially until a winner has been crowned. These competitions can be seen in video games, board games, endurance challenges, and reality competition series. We have developed a hierarchical Bayesian model to predict the results of a player-elimination tournament. In this talk we will use the reality competition series Project Runway as an illustrative example. We propose an adaptive model, which incorporates past results and player demographic information. We evaluate our model using various classification metrics.

Advancements in Transfer Learning for Health Outcomes (IS-57)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 111

Chair: Shane J Sacco

Proposer: Shane J Sacco, University of Connecticut Health Center

Presenters: Jun Jin, Xiaohui Yin, Shane J Sacco, Zheng Ren

Detection of Underdiagnosed Nonalcoholic Fatty Liver Disease in All of Us Cohort Using Prediction-Powered Inference with Automated Computational Phenotypes

Jun Jin, Henry Ford Health, Department of Public Health Sciences; Michigan State University, Department of Epidemiology and Biostatistics

Abstract. NAFLD, now termed MASLD and including NASH/MASH, is common yet underdiagnosed in primary care because liver biopsy and accurate noninvasive tests such as MRI-based assessment or FibroScan are not routinely performed. In 37,441 All of Us participants undergoing cardiometabolic monitoring, only 8,150 had definitive evaluation before July 1, 2021, including 2,375 positives; 473 additional patients evaluated by December 31, 2021 were used for out-of-sample testing. Obesity and diabetes differed significantly between evaluated and unevaluated groups (both $p<0.001$), confirming covariate shift. We compared labeled-only GLM, weighted GLM (wGLM) for adjusting covariate shift, and the prediction-powered inference with automated computational phenotypes (PPI-ACP), which addresses covariate shift and uses an external cirrhosis-based score correlated with NAFLD without requiring unbiasedness for the target outcome. PPI-ACP outperformed both GLM and wGLM, achieving AUC 0.82 (95% CI 0.79-0.86) and PPV/prevalence ratio 2.35 (2.17-2.50) at 0.90 specificity, with narrower coefficient CIs, and estimated full-cohort prevalence at 33.0% (31.9%-34.0%) versus 7.0% observed, indicating substantial underdiagnosis.

The Blessing of Multiple Outcomes: Modeling with Incomplete Essential Covariates in Electronic Health Records

Xiaohui Yin, University of Connecticut

Abstract. The increasing reliance on Electronic Health Records (EHR) for comprehensive health analysis and disease prediction is accompanied by significant challenges, particularly when some critical information, such as patient demographics (race, gender, age, etc.), is incomplete or inaccurate. Such deficiencies can compromise the integrity of health analysis, leading to biased estimates, inadequate mitigation or evaluation of health disparities, and underperforming prediction models. Conventional imputation techniques, ranging from simple summary statistic replacements to complex model-based methods, often fail to overcome these issues, especially when faced with large-scale datasets, a high rate of missingness, and instances of data missing not at random. Motivated by a suicide risk study with substantial missing data on demographics, we introduce an integrative learning framework that exploits the multivariate structure of health outcomes in EHRs to simultaneously impute missing covariates and fit prediction models.

The interrelations across different health outcomes, viewed as a blessing rather than a hindrance, enable the simultaneous imputation of the missing covariates, the estimation of their effects, and the refinement of prediction models for multivariate outcomes.

Improving Prediction of 30-day Pneumonia Readmissions using Data Fusion and Survey Data

Shane J Sacco, UConn Health Center

Abstract. Background: Published statistical models predicting hospital readmissions following inpatient admissions for pneumonia perform modestly (area under the receiver-operating characteristic curve [AUC-ROC]=0.60-0.70). This is possibly due to limited psychosocial information in medical records from which these models are typically built. If models are to be utilized to help prevent hospital readmissions, they must be improved. Methods: We developed models predicting 30-day hospital readmissions following pneumonia admissions among patients seen at an academic medical center (N=2,752). Using targeted data fusion, a transfer learning technique, we generated new psychosocial variables from participants in a large survey-based study, the Health and Retirement Study (N=11,390). We compared models using "conventional" features from medical records with models containing these features and generated psychosocial features. Results: The conventional model performed similar to published models, having an average AUC-ROC of 0.63 (95% CI: 0.61-0.65). Inclusion of psychosocial variables improved AUC-ROC to 0.68 (0.65-0.71). Sensitivities and positive predictive values also improved across the risk distribution. Discussion: By incorporating generated psychosocial variables from an external source, model performance was meaningfully improved. This study serves as a first step in understanding how fusion may improve identification of high-risk patients using models and may help enhance efforts to prevent unnecessary hospital readmissions.

Uncertainty-Aware Deep Learning for Significant Wave Height Forecasting using Attention-Enhanced LSTM Networks

Zheng Ren, University of Connecticut

Abstract. Reliable short-term significant wave height forecasting is essential for maritime safety, offshore operations, and coastal risk management, yet its performance is often constrained by uncertainties in wind forecasts. This study presents an attention-enhanced Long Short-Term Memory (LSTM) framework that explicitly accounts for wind forecast uncertainty through noise-augmented training and probabilistic modeling strategies. Wind and wave observations from the Western Long Island Sound (WLIS) buoy, together with operational North American Mesoscale (NAM) wind forecasts at Sikorsky Station, are used to quantify wind forecast errors and embed their statistical characteristics into model development. Three modeling strategies are explored: (1) a perfect-future model trained with idealized future wind inputs, (2) a noise-augmented model incorporating stochastic perturbations during training to improve robustness, and (3) a probabilistic model that jointly forecasts wave height and associated uncertainty. The noise-augmented strategy reduces RMSE and MAE by roughly 7% relative to the perfect-future model and demonstrates improved stability across forecast lead times under varying wind uncertainty conditions. The probabilistic model achieves similar predictive accuracy while providing reliable uncertainty quantification. Both strategies consistently outperform a baseline relying solely on historical buoy measurements. These results highlight the importance of explicitly addressing wind forecast uncertainty in data-driven wave forecasting models. The proposed framework enhances both accuracy and reliability, offering a practical solution for operational wave forecasting in the presence of imperfect atmospheric inputs and supporting more informed coastal and offshore decision-making.

Sufficient Dimension Reduction and Machine Learning for Complex Data (IS-59)

Time & Location: May 28, 08:30 AM - 10:10 AM | Room 302

Chair: Hossein Moradi Rekabdarkolaee

Proposer: Shanshan Ding, Chenlu Ke, University of Delaware, Virginia Commonwealth University

Presenters: Shanshan Ding, Chenlu Ke, Hossein Moradi Rekabdarkolaee

Integrative dimension reduction for high dimensional multi-source data

Shanshan Ding, University of Delaware

Abstract. In this talk, an integrative sufficient dimension reduction method is proposed to achieve simultaneous dimension reduction and variable selection for multi-source data analysis in high dimensions. The proposed method aims to extract sufficient information in a supervised fashion, and the asymptotic results establish a new theory for integrative dimension reduction. The promising performance of the integrative estimator and efficient numerical algorithms is demonstrated through simulation and real data analysis.

Explained Kernel Embedding Variation for Sufficient Dimension Reduction with Extensions to Censored Outcomes

Chenlu Ke, Virginia Commonwealth University

Abstract. We propose a new framework for sufficient dimension reduction based on explained kernel embedding variation. The main idea is to view dimension reduction through a model-free ANOVA decomposition in reproducing kernel Hilbert spaces and to estimate low-dimensional linear combinations of predictors that maximize kernel regression sum of squares. This formulation unifies several SDR targets: with characteristic kernels it targets the central subspace, while with linear kernels it reduces to central mean subspace estimation. On the computational side, we combine a sequential procedure that builds an ordered set of directions with a joint refinement step over the Grassmann manifold. We also use a validation-based backward test to determine the structural dimension. We further extend the framework to right-censored outcomes and develop an IPCW version of kernel regression sum of squares for estimating sufficient reductions for the event time. Numerical studies show that the proposed method can recover the relevant dimension and directions in both fully observed and censored settings.

Computer Vision in Action: Pinkeye Detection in Cattle

Hossein Moradi Rekabdarkolaee, Bowling Green State University

Abstract. Bovine pinkeye is a contagious ocular disease that significantly impacts cattle health and agricultural productivity. Conventional diagnostic methods rely on clinical observation, which can be subjective, time-consuming, and impractical for large herds or remote locations. To address these limitations, this study explores the application of computer vision algorithms for automated pinkeye detection using Deep Learning. Furthermore, we employed different generative AI algorithms to address the imbalance data. Our findings shows the effectiveness of the proposed approach.

Parallel Session 2 | 02:00 PM - 03:40 PM, May 28

Real-world applications of statistical analysis for complex data (IS-6)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 101

Chair: Hyemin Yeon

Proposer: Hyemin Yeon, Kent State University

Presenters: James Stephen Marron, David Hitchcock, Dan Kowal, Margaret Hoch

Concurrent Functional Regression to Reconstruct River Stage Data during Flood Events

David B. Hitchcock, University of South Carolina

Abstract. On October 4, 2015, the Cedar Creek gage at Congaree National Park stopped reporting stages, and the readings did not resume until approximately 2 weeks later because of record-breaking rainfall that led to some of the worst floodings in South Carolina history. Our goal is to reconstruct the Cedar Creek stage during this missing 2-week window. Our analysis uses a sample of ten historical flood events from the last 25 years. The Congaree River gage in Congaree National Park remained functioning throughout the October 2015 flood, when the stage reached its maximum recorded crest. The stages from the two gages are directly related during floods. We introduce a new method to objectively determine the start and end points of each flood event in the sample and then use these events to predict the missing Cedar Creek stage. We treat the stage as functional data and use a concurrent model to establish the relationship between the two locations during each timepoint of prior flood events. Once this relationship is found, the known Congaree stage is used to predict the missing Cedar Creek stage during the 2015 flood. The results show that there is a strong functional relationship between the two locations, and that the crest of Cedar Creek is a historic high, reaching stages above 17 feet, with a previous high of just over 16 feet. We also include a new stage prediction using our model for a recent flood event during

Hurricane Helene in September 2024, and the model predicts the actual Cedar Creek stage very well.

Bayesian regression with fragmented and noisy functional covariates

Dan Kowal, Cornell University

Abstract. Sparsely-observed and noisy functional data appear widely, especially in longitudinal data and in wearable device data where non-wear periods result in highly fragmented observations. Modeling such data requires great care to avoid overfitting and underestimation of uncertainty. Additional challenges arise when sparsely-observed and noisy functions appear as covariates in a regression model. In that case, poorly-estimated curves may attenuate or suppress the underlying regression effects, while failure to account for the substantial uncertainties in these functional covariates can corrupt inference. We propose a fully Bayesian scalar-on-function regression model for sparsely-observed functional covariates measured with error. This approach features a functional factor model to smooth, impute, and borrow information across fragmented curve observations. Crucially, our MCMC algorithm deviates from predominant existing approaches in that it jointly samples the posterior and imputes the functional covariates together in one block, leading to efficiency and stability under high missingness. Using simulated data, we demonstrate the substantial limitations of existing methods with sparsely-observed, fragmented, and noisy functional covariates, and show how the proposed approach maintains accurate estimation and precise, well-calibrated inference. We apply our functional factor and scalar-on-function regression models to a bone mineral density longitudinal study and wearable device physical activity data.

Linear Regression Using Principal Components from General Hilbert-Space-Valued Covariates

Margaret Hoch, University of North Carolina Chapel Hill

Abstract. We introduce Adaptive Subspace PCA (AS-PCA), a framework for principal component analysis of random elements in a general separable Hilbert space. AS-PCA projects the covariance operator onto a data-adaptive finite-dimensional subspace prior to eigendecomposition, requiring no kernel specification and accommodating multi-dimensional functional objects including images and surfaces. Under the second-moment condition, we prove a Donsker theorem for Hilbert-space-valued empirical processes and use it to establish uniform consistency and joint Gaussian limits for the leading eigenpairs. A data-driven diagnostic verifies projection accuracy, and a consistent proportion-of-variance-explained rule selects the number of components. Building on AS-PCA, we construct Hilbert-Space Principal Component Regression (HS-PCR) for models combining Euclidean and Hilbert-space-valued covariates. The HS-PCR estimator is root-$n$consistent and asymptotically normal, with an explicit influence function decomposition accounting for eigenfunction estimation uncertainty. Both nonparametric and wild bootstrap procedures are shown to be asymptotically valid. Simulations with two- and three-dimensional imaging predictors confirm accurate eigenstructure recovery and nominal bootstrap coverage. HS-PCR is applied to Alzheimer's Disease Neuroimaging Initiative data in regression and precision-medicine settings.

Innovative Statistical Designs and Methods for Efficient and Robust Clinical Trials (IS-8)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 110

Chair: Xiaohan Guo

Proposer: Haitao Chu, Pfizer Inc.

Presenters: Yunqi Zhao, Fan Li, Xiaohan Guo, Vincent Tan

BEAM: Bayesian Hybrid Design With Adaptive Sample Size Through Multisource Exchangeability Modeling

Yunqi Zhao, Takeda

Abstract. Randomized controlled trials (RCTs) are considered the gold standard for evaluating treatment efficacy, but they come with several practical challenges. These include high costs, lengthy timelines, ethical concerns for participants in placebo or control arms, and issues such as patient attrition and non-compliance. Recruiting patients for the control arm can be particularly challenging, especially in therapeutic areas with high unmet medical needs. To address these issues, hybrid trial designs that integrate external data sources, such as historical controls and real-world data, have emerged as a promising alternative. This paper introduces the Bayesian hybrid design with adaptive

sample size through multisource exchangeability modeling (BEAM). The BEAM design leverages a modified multisource exchangeability model to dynamically borrow relevant information from multiple historical data sources, while adaptively adjusting the sample size throughout the trial. This approach ensures that the trial maintains statistical rigor and efficiency, even when heterogeneity exists between current and historical data, and mitigates the challenges associated with control arm accrual and compliance. Through extensive simulations, BEAM demonstrated robust performance in controlling type I error rate, reducing bias, and maintaining power compared to traditional methods and other adaptive designs. Additionally, the BEAM design offers a versatile and efficient computational framework for optimizing clinical trials, helping to reduce both the cost and time involved in drug development. We also illustrate the application of the proposed BEAM design in a case study on ankylosing spondylitis.

Optimal estimation of generalized causal effects in cluster-randomized trials with multiple outcomes

Fan Li, Yale School of Public Health

Abstract. Cluster-randomized trials (CRTs) are widely used to evaluate group-level interventions and increasingly collect multiple outcomes capturing complementary dimensions of benefit and risk. Investigators often seek a single global summary of treatment effect, yet existing methods largely focus on single-outcome estimands or rely on model-based procedures with unclear causal interpretation or limited robustness. We develop a unified potential outcomes framework for generalized treatment effects with multiple outcomes in CRTs, accommodating both non-prioritized and prioritized outcome settings. The proposed cluster-pair and individual-pair causal estimands are defined through flexible pairwise contrast functions and explicitly account for potentially informative cluster sizes. We establish nonparametric estimation via weighted clustered U-statistics and derive efficient influence functions to construct covariate-adjusted estimators that integrate debiased machine learning with U-statistics. The resulting estimators are consistent and asymptotically normal, attain the semiparametric efficiency bounds under mild regularity conditions, and have analytically tractable variance estimators that are proven to be consistent under cross-fitting. Simulations and an application to a CRT for chronic pain management illustrate the practical utility of the proposed methods.

Adaptive Seamless Phase 2/3 Designs with Contribution of Component Demonstration in Oncology Trails

Xiaohan Guo, Pfizer

Abstract. Combination therapy is a cornerstone of anticancer drug development. In addition to demonstrating a significant treatment effect, establishing the contribution of each individual agent to the overall clinical activity is critical to the success of a combination therapy. Recent FDA guidance recommends randomized trials to demonstrate the Contribution of Components (COC). When retrospective analyses using data from early-phase trials or historical sources are insufficient for COC evaluation, sponsors may need to conduct a multi-arm phase 3 trial that includes randomized monotherapy and combination therapy arms, or an additional randomized phase 2 trial specifically designed for COC assessment. However, both approaches can substantially delay the development of new combination therapies. To address these challenges, we propose an adaptive seamless phase 2/3 design. In this design, Stage 1 employs a multi-arm approach that includes randomized monotherapy and combination therapy arms, enabling COC evaluation and informing treatment selection. Stage 2 then focuses on testing the treatment effect of the selected regimen. The family-wise type I error rate is controlled through appropriate multiplicity adjustments to account for the multi-stage, multi-arm structure. The overall sample size is determined based on the joint requirements for demonstrating treatment effect and establishing COC. Simulation studies are conducted to compare the operating characteristics of several design options within this framework. A case study and practical recommendations are also provided to illustrate the implementation of the proposed design in real-world oncology trials.

Multiple Imputation for Small, Extremely High Efficacy Clinical Trials with Binary Endpoints

Yaoyuan Vincent Tan, Vertex Pharmaceuticals

Abstract. There has been an increasing interest in using cell and gene therapy (CGT) to treat/cure difficult diseases. The hallmark of CGT trials are the small sample size and extremely high efficacy. Due to the innovation and novelty of such therapies, when there is missing data, more scrutiny is exercised, and regulators often request for missing data

handling strategy when missing data occurs. Often, multiple imputation (MI) will be used. MI for continuous endpoint is well established but literature of MI for binary endpoint is lacking. In this work, we compare and develop 3 new methods to handle missing data using MI for binary endpoints when the sample size is small and efficacy extremely high. The parameter of interest is population proportion of success. We show that our proposed methods performed well and produced good 95% coverage. We also applied our methods to an actual clinical study, the Clinical Islet Transplantation (CIT) Protocol 07, conducted by National Institutes of Health (NIH).

Modern Econometric Methods for Complex Economic and Financial Systems (IS-16)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 305

Chair: Haim Bar

Proposer: Haim Bar, University of Connecticut

Presenters: Jackson P. Lautier, Martin T. Wells, Thomas F. P. Wiesen

Discrete Time-to-Event Regression Analysis Under Left-Truncation with Applications to Consumer Finance

Jackson P. Lautier, Bentley University

Abstract. Asset-backed securities (ABS) play a vital role in financing American consumer automobile debt. Recently, economic analysis into ABS has benefited from the public release of data, which provides a new, rich source of loan level consumer auto loan information. Because this loan lifetime data is discrete-time and subject to random left-truncation, however, it is nontrivial to analyze. This has attracted recent study, but there is still no suitable approach to model this ABS loan lifetime data that incorporates regression coefficients for the lifetime of interest. We thus generalize the conditional, bivariate distribution to link to covariates, while keeping the left-truncation distribution unspecified. We solve the high-dimensional, constrained likelihood-based parameter estimation problem numerically, using a block coordinate descent design. Under suitable regularity conditions, we provide the complete large sample, multivariate normal distribution of the estimators. This allows for large sample inference into variable and model selection. We both prove all results and verify them through simulation studies. Our methods are then applied in an economic study of borrower prepayment behavior for 1,553 consumer auto loans from the 2017-3 Ally Auto Receivables Trust ABS bond. We find that borrowers with pick-up trucks prepay slower, all else equal, among other consumer finance insights.

Is There an AI Bubble? Robust Date-Stamping for Periods of Exuberance

Martin Wells, Cornell University

Abstract. The recent surge in valuations of firms exposed to artificial intelligence has raised concerns about speculative exuberance in technology and semiconductor markets. Motivated by this episode, we develop a stochastic-volatility-robust ADF (SV-ADF) framework for detecting and date-stamping asset-price bubbles when prices exhibit persistent, time-varying volatility. Standard right-tailed Dickey–Fuller procedures can be distorted in heteroskedastic environments, leading to unstable inference and spurious bubble classifications. We extend recursive ADF testing to autoregressive models with highly persistent mean and volatility components and derive nuisance-parameter-free calibration rules for both bubble origination and collapse. The resulting procedure delivers consistent date-stamping, separates origination and collapse thresholds, and remains implementable for real-time monitoring. Monte Carlo evidence shows improved power and dating accuracy relative to homoskedastic PWY tests, especially under pronounced volatility persistence. Applying the method to AI-exposed equities, including the Magnificent Seven and major semiconductor firms, we find substantial cross-sectional heterogeneity in exuberance, with current bubble dynamics concentrated in Alphabet and TSMC, while Tesla and Nvidia exhibited earlier explosive episodes.

Urban Influence of America's Largest Metropolitan Areas on US Counties: Evidence from Econometric Connectedness

Thomas F. P. Wiesen, University of Maine

Abstract. The objective of this paper is to examine the econometric connectedness of US counties relative to the country's largest metropolitan areas. The motivating background is an interest in measuring the urban influence of US counties to compare the economic performance of regions across the urban-to-rural spectrum and examine the effects of "rurality." To measure urban influence, we utilize

the concept of econometric connectedness which is commonly used to measure market integration in finance and macroeconomics. Using monthly employment figures for all counties from 1990 to the present, separate vector autoregressive (VAR) models are estimated, which include New York, Los Angeles, Chicago, a county of interest, the US overall, and the state where the county of interest is located. These employment figures serve as a proxy for a region's overall economic activity. We then use joint conditioning sets and the novel joint forecast error variance decomposition to extract the percent of the forecast error variance of a US county's economic activity explained by the shocks of the metropolitan areas included in the model. This measure of econometric connectedness quantifies how influential the metropolitan areas are in explaining the economic activity of US counties, and it generates a continuous urban influence score between zero and 100% for each county. A second version of the analysis measures urban influence relative to the nine largest US metropolitan areas, which adds Dallas, Houston, Washington DC, Philadelphia, Miami, and Atlanta. The econometric connectedness results show a wide variation in the urban influence of US counties ranging from places with urban influence scores of close to zero to counties with scores approaching 90%. The paper's main policy implication is that it provides a new measure of urban influence to complement existing indicators such as the urban influence codes of the US Department of Agriculture (USDA), which places counties in one of twelve discrete categories based on their population size and proximity to an urban area. Unlike the USDA urban influence codes, our measure of econometric connectedness has a continuous scale, and it is based on the economic integration of US counties to metropolitan areas.

Advancing Machine Learning, AI and Frontier Methods for Clinical Decision Making and Trial Design (IS-18)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 205

Chair: Zhaowei (Zoe) Hua

Proposer: Zhaoyang Teng, Astellas Pharma

Presenters: Yukang Jiang, Yusuke Yamaguchi, Qingkai Dong, Evelyn Zheng

Real-World Evidence from Electronic Health Records: Foundation AI Models for Clinical Decision Support

Yukang Jiang, The University of North Carolina at Chapel Hill

Abstract. Real-world data from electronic health records are becoming an important foundation for clinical decision support. Unlike traditional research datasets, EHR data are longitudinal, heterogeneous, irregularly sampled, and often affected by informative missingness. These properties make simple feature-vector models insufficient for many clinical prediction and phenotyping tasks. In this talk, I will discuss how recent AI methods, especially EHR foundation models, aim to learn reusable patient representations from routine care data. I will introduce major modeling directions, including structured-sequence models, next-event prediction, time-to-event foundation models, generative patient trajectory models, and multimodal fusion with clinical notes, imaging, genomics, and other omics data. I will also discuss how these approaches can support risk prediction, phenotyping, subtyping, disease trajectory modeling, and clinical decision support. Finally, I will highlight remaining challenges, including reliability, generalization across hospitals and populations, noisy proxy labels, privacy, calibration, and safe clinical deployment. The overall goal is to move from static EHR-based prediction toward trustworthy, longitudinal, and clinically useful AI systems for real-world evidence generation and patient-centered decision support.

Propensity score-based unequal matching for rare disease clinical trials with external controls

Yusuke Yamaguchi, Astellas

Abstract. Externally controlled trial designs are increasingly used in rare disease drug development when randomized controlled trials are infeasible due to limited patient populations or ethical constraints. Propensity score (PS) matching is commonly applied to improve comparability between trial participants and external controls; however, existing methods are primarily designed for settings where the number of matched external controls is at least as large as the number of treated patients. In rare diseases and hybrid control designs, investigators may instead require fewer matched external controls than trial patients, creating a reversed-ratio setting that we refer to as unequal matching. We formulate unequal matching as a constrained subset-selection

problem and present a distributional matching method using a genetic algorithm, which can be categorized as a global optimization method directly optimizing agreement between the overall PS distributions of trial and matched external control populations. Simulation studies suggested that the method consistently achieved superior PS balance, reduced bias, and maintained type I error rates near the nominal level. Unequal PS matching provides a practical and robust framework for externally controlled rare disease trials.

Predictive-Modeling–Assisted Interim Decision-Making in Adaptive Trials with Censored Survival Outcomes

Qingkai Dong, University of Connecticut

Abstract. Adaptive designs for studies with censored survival endpoints rely on interim analyses to guide decisions such as early stopping for futility, yet these decisions are commonly driven by conditional power estimates that extrapolate interim treatment effects and ignore prognostic baseline information. For log-rank–based analyses, this practice can lead to inaccurate forecasts of final results, particularly when censoring is substantial. We develop a predictive-modeling–assisted framework that incorporates baseline covariates at the interim stage to improve estimation of post-interim treatment effects and conditional power while preserving standard log-rank–based decision rules. The approach augments interim data by predicting event times for censored participants and uses the resulting completed dataset solely for interim decision support. We introduce evaluation metrics that assess both conditional power accuracy and the quality of interim futility decisions. Extensive simulation studies characterize when the proposed framework improves or degrades performance, highlighting the roles of covariate informativeness, covariate type, and prediction range. The results provide practical guidance for using predictive models to support interim decision-making in studies with time-to-event outcomes.

Adaptive Design with Interim Analysis for Enrichment and Sample Size Re-estimation

Evelyn Zheng, Vertex Pharmaceuticals

Abstract. Identifying the appropriate target population for Phase 3 clinical trials is a key challenge in clinical development. While a biomarker may exist that predicts treatment response, available data are often limited for reliably determining an optimal cutoff. When the trial population is enriched based on interim findings, an increase in sample size may be required to maintain adequate power, particularly if many subjects enrolled before enrichment do not belong to the targeted subgroup. We propose an adaptive design incorporating interim analysis for biomarker-based enrichment and sample size re-estimation. The adaptive rule is guided by the concept of a "promising zone" for conditional power, considering both the overall and subgroup-specific treatment effects. We evaluate the operating characteristics through simulation in a two-arm placebo controlled randomized trial with a continuous outcome and a predictive biomarker. Simulation results demonstrate improved power with the adaptive design across a range of scenarios reflecting different biomarker-treatment relationships, as well as preserved overall type I error rate. The proposed design extends existing adaptive methodologies by jointly incorporating enrichment and sample size re-estimation. It offers a practical approach when a predictive biomarker is available, but a meaningful cutoff remains uncertain. With appropriate consideration of biomarker performance, interim timing, and operational feasibility, this design can enhance trial efficiency, uphold ethical standards, and ensure adequate statistical power.

Causal analysis and beyond in the era of modern data science (IS-21)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 206

Chair: Riyanka Bhowal, Department of Statistics, University of Connecticut

Proposer: Dipak K Dey, UCONN

Presenters: Subharup Guha, Abhishek Chakrabortty, Nabarun Deb, Priyam Das.

Causal Meta-Analysis by Integrating Multiple Observational Studies with Multivariate Outcomes

Subharup Guha, Dartmouth College

Abstract. Integrating multiple observational studies to make unconfounded causal or descriptive comparisons of group potential outcomes in a large natural population is challenging. Moreover, retrospective cohorts, being convenience samples, are usually unrepresentative of the natural population of interest and have groups with unbalanced covariates.

We propose a general covariate-balancing framework based on pseudo-populations that extends established weighting methods to the meta-analysis of multiple retrospective cohorts with multiple groups. Additionally, by maximizing the effective sample sizes of the cohorts, we propose a FLEXible, Optimized, and Realistic (FLEXOR) weighting method appropriate for integrative analyses. We develop new weighted estimators for unconfounded inferences on wide-ranging population-level features and estimands relevant to group comparisons of quantitative, categorical, or multivariate outcomes. Asymptotic properties of these estimators are examined. Through simulation studies and meta-analyses of TCGA datasets, we demonstrate the versatility and reliability of the proposed weighting strategy, especially for the FLEXOR pseudo-population.

Doubly Robust Causal Inference with Partially Labeled Data: The Decaying MAR Framework

Abhishek Chakrabortty, Texas A&M University

Abstract. In modern large-scale observational studies, data collection constraints often result in partially labeled datasets, posing challenges for reliable causal inference, especially due to potential labeling bias and relatively small size of the labeled data. This paper introduces a decaying missing-at-random (decaying MAR) framework and associated approaches for doubly robust causal inference on treatment effects in such semi-supervised (SS) settings. This simultaneously addresses selection bias in the labeling mechanism and the extreme imbalance between labeled and unlabeled groups, bridging the gap between the standard SS and missing data literatures, while throughout allowing for confounded treatment assignment and high-dimensional con-founders under appropriate sparsity conditions. To ensure robust causal conclusions, we propose a bias-reduced SS (BRSS) estimator for the average treatment effect, a type of 'model doubly robust' estimator appropriate for such settings, establishing asymptotic normality at the appropriate rate under decaying labeling propensity scores, provided that at least one nuisance model is correctly specified. Our approach also relaxes sparsity conditions beyond those required in existing methods, including standard supervised approaches. Numerical experiments confirm the effectiveness and adaptability of our estimators in addressing labeling bias and model misspecification.

Robustness and Efficiency of Rosenbaum's Rank-based Estimator in Randomized Trials: A Design-based Perspective

Nabarun Deb, University of Chicago

Abstract. Mean-based estimators of causal effects in randomized experiments may behave poorly if the potential outcomes have a heavy tail or contain outliers. An alternative estimator proposed by Rosenbaum (1993) estimates a constant additive treatment effect by inverting a randomization test using ranks. We develop a design-based asymptotic theory for this rank-based estimator and study its robustness and efficiency properties. We show that Rosenbaum's estimator is robust against outliers with a breakdown point that uniformly dominates that of any weighted quantile estimator. When pretreatment covariates are available, a regression-adjusted version of Rosenbaum's estimator uses an agnostic linear regression on the covariates and bases inference on the ranks of residuals. Under mild integrability conditions, we show that this estimator is at most 13.6% less efficient, in the worst case, than the commonly used mean-based regression adjustment method proposed by Lin (2013); often outperforming it when the residuals have heavy tails. Moreover, under suitable assumptions, Rosenbaum's regression-adjusted estimator is at least as efficient as the unadjusted one. Finally, we initiate the study of Rosenbaum's estimator when the constant treatment effect assumption may be violated. To analyze the regression-adjusted estimator, we develop local asymptotics of rank statistics under the design-based framework, which may be of independent interest.

SMART-MC: Characterizing the Dynamics of Multiple Sclerosis Therapy Transitions Using a Covariate-Based Markov Model

Priyam Das, Virginia Commonwealth University

Abstract. Treatment switching is a common occurrence in the management of Multiple Sclerosis (MS), where patients transition across various disease-modifying therapies (DMTs) due to heterogeneous treatment responses, differences in disease progression, patient characteristics, and therapy-associated adverse effects. To investigate how patient-level covariates influence the likelihood of treatment transitions among DMTs, we adopt a Markovian framework, Sparse Matrix Estimation with Covariate-Based Transitions in Markov Chain Modeling (SMART-MC), in which the transition probabilities are modeled as functions of these co-

variates. Modeling real-world treatment transitions under this framework presents several challenges, including ensuring parameter identifiability and handling sparse transitions without overfitting. To address identifiability, we constrain each transition-specific covariate coefficient vectors to have a fixed L2 norm. Furthermore, our method automatically estimates transition probabilities for sparsely observed transitions as constants and enforces zero transition probabilities for transitions that are empirically unobserved. This approach mitigates the need for additional model complexity to handle sparsity while maintaining interpretability and efficiency. To optimize the multi-modal likelihood function, we develop a scalable, parallelized global optimization routine, which is validated through benchmark comparisons and supported by key theoretical properties. Our analysis uncovers meaningful patterns in DMT transitions, revealing variations across MS patient subgroups defined by age, race, and other clinical factors. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

Robust Causal Inference with Post-Treatment Events and Survival Outcomes (IS-28)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 302

Chair: Guangyu Tong

Proposer: Guangyu Tong, Yale University

Presenters: Chao Cheng, Sean McGrath, Guanbo Wang, Henan Xu

Identification and Multiply Robust Estimation in Causal Mediation Analysis Across Principal Strata

Chao Cheng, Washington University in St. Louis

Abstract. We consider assessing causal mediation in the presence of a post-treatment event (examples include noncompliance, a clinical event, or death). We identify natural mediation effects for the entire study population and for each principal stratum characterized by the joint potential values of the post-treatment event. We derive the efficient influence function for each mediation estimand, which motivates a set of multiply robust estimators for inference. The multiply robust estimators are consistent under four types of misspecifications and are efficient when all nuisance models are correctly specified. We also develop a nonparametric efficient estimator that leverages data-adaptive machine learners to achieve efficient inference and discuss sensitivity methods to address key identification assumptions. We illustrate our methods via simulations and two real data examples.

Time-smoothed inverse probability weighted estimation of effects of generalized time-varying treatment strategies on repeated outcomes truncated by death

Sean McGrath, Yale School of Public Health

Abstract. Researchers are often interested in estimating effects of generalized time-varying treatment strategies on the mean of an outcome at one or more selected follow-up times of interest. For example, the Medications and Weight Gain in PCORnet (MedWeight) study aimed to estimate effects of adhering to flexible medication regimes on future weight change using electronic health records (EHR) data. This problem presents several methodological challenges that have not been jointly addressed in the prior literature. First, this setting involves treatment strategies that vary over time and depend dynamically and non-deterministically on measured confounder history. Second, the outcome is repeatedly, non-monotonically, informatively, and sparsely measured in the data source. Third, some individuals die during follow-up, rendering the outcome of interest undefined at the follow-up time of interest. In this article, we pose a range of inverse probability weighted (IPW) estimators targeting effects of generalized time-varying treatment strategies in truncation by death settings that allow time-smoothing for precision gain. We conducted simulation studies that confirm precision gains of the time-smoothed IPW approaches over more conventional IPW approaches that do not leverage the repeated outcome measurements. We illustrate an application of the IPW approaches to estimate comparative effects of adhering to flexible antidepressant medication strategies on future weight change. The methods are implemented in the accompanying R package, smoothedIPW.

Safe Trial Augmentation for Survival Outcomes

Guanbo Wang, Dartmouth College

Abstract. Randomized trials often lack sufficient sample sizes to provide precise estimates of average treatment effects, particularly when outcomes are

time-to-event and subject to censoring. Incorporating external data is a promising strategy to improve statistical efficiency, but incompatibility between external and trial data can introduce bias in estimating effects for the trial population. Recently, Wang et al. proposed a method that leverages external data to improve efficiency while preserving consistency, even when incompatibility is present. However, this approach does not accommodate survival outcomes. In this work, we extend the framework to develop a robust augmentation procedure for time-to-event endpoints. Our method guarantees that, whenever trial-only analyzes yield consistent estimates of the ATE, our estimator is also consistent and never performs worse—in terms of asymptotic efficiency—than the most efficient estimator based solely on trial data.

Marginal functional mediation analysis with zero-inflated counts

Henan Xu, Yale School of Public Health

Abstract. Zero-inflated count outcomes commonly arise in biomedical studies, and in many such settings, investigators seek to decompose treatment effects into pathway-specific components operating through a time-varying mediator observed sparsely and irregularly. In the Wisconsin Smokers' Health Study 2 (WSHS2), for example, investigators may be interested in how the effect of varenicline on subsequent cigarette use is mediated by time-varying craving measured during the ecological momentary assessment period. Motivated by this study, we propose a marginal causal mediation framework for a binary treatment, a sparse functional mediator, and a zero-inflated count outcome by combining a functional regression model for the mediator process with a functional marginalised zero-inflated Poisson model for the outcome. Within the potential outcomes framework, we define natural direct and indirect effects on the incidence-rate-ratio and risk-difference scales and show that the causal estimands admit closed-form expressions, enabling interpretable summaries of how mediation accumulates over time. We develop estimators and inference procedures for the causal effects, establish their asymptotic properties, and provide a bias-formula sensitivity analysis for unmeasured mediator–outcome confounding. Simulation studies evaluate finite-sample performance, and we return to the motivating WSHS2 data to quantify the extent to which craving mediates the effect of varenicline on later cigarette use.

Recent Advances in Analysis of Network Data (IS-43)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 201

Chair: Wenrui Li

Proposer: Wenrui Li, University of Connecticut

Presenters: Ying Guo, Lizhen Lin, Michael Schweinberger, Johan Ugander

A Statistical AI Framework for Learning Directed Brain Connectomes from Neuroimaging Data

Ying Guo, Emory University

Abstract. In recent years, connectome-based research has become a central focus in neuroscience, offering essential insights into brain organization and advancing predictive modeling of cognitive, behavioral, and mental health outcomes. While most existing approaches focus on undirected brain connectivity, they overlook the directionality and causal influences between brain regions. To address this limitation, we propose a statistical AI method for learning directed brain connectomes from neuroimaging data. Our approach integrates principled statistical modeling with deep learning to infer sparse, interpretable directed connectivity graphs that characterize latent causal interactions across the brain. At the same time, the method learns low-dimensional graph embeddings optimized for downstream prediction tasks, including demographic attributes and clinical phenotypes. The proposed method uncovers whole-brain directed connectivity patterns and reveals novel subpopulation-specific connectomic differences, highlighting its potential to advance both mechanistic understanding and predictive modeling in neuroscience.

Exchangeable random permutations with an application to Bayesian graph matching

Lizhen Lin, The University of Maryland

Abstract. We introduce a general Bayesian framework for graph matching grounded in a new theory of *exchangeable random permutations*. Leveraging the cycle representation of permutations and the literature on exchangeable random partitions, we define, characterize, and study the structural and predictive properties of these probabilistic objects. A

novel sequential metaphor, the *position-aware generalized Chinese restaurant process*, provides a constructive foundation for this theory and supports practical algorithmic design. Exchangeable random permutations offer flexible priors for a wide range of inferential problems centered on permutations. As an application, we develop a Bayesian model for graph matching that integrates a correlated stochastic block model with our novel class of priors. The cycle structure of the matching is linked to latent node partitions that explain connectivity patterns, an assumption consistent with the homogeneity requirement underlying the graph matching task itself. Posterior inference is performed through a node-wise blocked Gibbs sampler directly enabled by the proposed sequential construction. To summarize posterior uncertainty, we introduce *perSALSO*, an adaptation of SALSO to the permutation domain that provides principled point estimation and interpretable posterior summaries. Together, these contributions establish a unified probabilistic framework for modeling, inference, and uncertainty quantification over permutations.

Causal inference in connected populations with contagion

Michael Schweinberger, Department of Statistics, The Pennsylvania State University

Abstract. Causal inference in connected populations is complicated by contagion and other real-world processes inducing complex dependence among outcomes. While there is a growing body of work on estimating causal effects under contagion, little is known about how correlations among outcomes induced by contagion impact causal effects and inference. We provide insight into how contagion affects interventions based on closed-form expressions for causal effects under contagion, which are the first such results to our knowledge. These closed-form expressions demonstrate that the effects of interventions are intertwined even in the simplest possible settings, and that contagion can decrease or increase the effects of interventions. We discuss statistical implications, including violations of neighborhood exposure assumptions by contagion, asymptotic bias of model-based estimators of causal effects ignoring correlations among outcomes due to contagion, and design-based estimators of causal effects.

Causal inference under structured interference

Johan Ugander, Yale University

Abstract. The field of causal inference develops methods for estimating treatment effects, often relying on the Stable Unit Treatment Value Assumption (SUTVA), which states that a unit's outcome depends only on its own treatment. However, in many real-world settings, SUTVA is violated due to interference, where the treatment assigned to one unit influences the outcomes of others. Interference can arise from social interactions, competition for shared resources, or other forms of coupling, all of which complicate causal analysis and lead to biased estimates of treatment effects when SUTVA is assumed. In many cases, interference follows structured patterns that we argue can be leveraged for more accurate estimation. In this paper, we examine and formalize two specific forms of structured interference, monotone interference and submodular interference, which we argue arise in many practical settings. We investigate how incorporating these structures can improve causal effect estimation. Our main contributions are (i) a set of bounds relating key interference estimands under these structural assumptions and (ii) new estimators that integrate these structures through constrained optimization. Since these constraints may introduce bias, we further develop debiasing techniques based on treatment regeneration and bootstrap methods, which we find to be effective. Joint work with Kevin Han and Shuangning Li.

Methodological Advances for Biomedical Data: Reinforcement Learning, Graphical Models, and Causal Inference (IS-46)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 301

Chair: Shuangge Ma

Proposer: Shuangge Ma, Yale University

Presenters: Pangpang Liu, Ruyi Liu, Mingcong Wu, Jingmao Li

Dynamic Feature Acquisition for Multi-Label Learning Using Reinforcement Learning

Pangpang Liu, Yale University

Abstract. In many real-world applications, acquiring features incurs cost, and the optimal feature subset can vary across instances. This challenge is amplified in multi-label learning, where label dependencies influence predictive performance. We propose Cost-Aware Feature Acquisition for Multi-Label learning

(CAFA-ML), a novel framework that learns a sequential feature acquisition policy and a unified multi-label predictor. CAFA-ML uses reinforcement learning (RL) to select informative feature sets within a budget, and employs a mask-based predictor with a multi-label graph to capture label dependence. Experiments on real datasets demonstrate that CAFA-ML consistently achieves superior predictive performance while reducing feature acquisition cost compared to state-of-the-art multi-label and RL-based baselines.

Bayesian inference for the causal hazard ratio

Ruyi Liu, Department of Biostatistics, Yale School of Public Health, New Haven, Connecticut, U.S.A.

Abstract. The hazard ratio is the most widely reported effect measure in clinical trials and observational studies with time-to-event outcomes, yet its causal interpretation is fundamentally compromised by a built-in selection bias: even in randomized trials, the at-risk populations under treatment and control differ systematically in frailty composition over time. A principled alternative is the causal hazard ratio, which restricts comparison to the group of individuals who would survive past any given time under either treatment assignment. The existing method identifies the effects through a parametric frailty model. However, this approach requires an approximation in observational settings, which is valid only under rare event rates or weak confounding, and fails to accommodate covariate adjustment in randomized trials. We propose a novel copula-based analysis framework that achieves identification of the causal hazard ratio without approximation in both randomized and observational settings. Our approach provides closed-form identification formulas for a flexible family of copulas and enables principled covariate adjustment via a conditional copula assumption. We further develop a Bayesian inference framework using outcome models that range from parametric accelerated failure time models to flexible Bayesian nonparametric alternatives, such as Bayesian additive regression trees for survival. This approach delivers coherent uncertainty quantification through posterior inference. Simulation studies demonstrate that our estimator achieves near-nominal credible interval coverage and small bias across data-generating mechanisms of varying complexity. In contrast, the existing estimator exhibits severe bias and markedly below-nominal coverage in the presence of covariate heterogeneity and nonlinear treatment-effect surfaces.

Assisted disease network analysis with a temporal conditional graphical model based on electronic health records.

Mingcong Wu, Yale School of Public Health

Abstract. Leveraging electronic health records (EHRs) to elucidate interconnections among diseases has become increasingly vital in advancing healthcare, with disease network analysis taking a central role in uncovering intricate relationships. In this study, we introduce an assisted disease network analysis framework. By integrating the gene expression data, our approach enhances the disease network with molecular context, offering a more comprehensive perspective on disease-gene interactions. Moreover, most existing network studies fall short in capturing the temporal relationships embedded in longitudinal records. To fill this gap, we propose a temporal conditional graphical model designed to simultaneously account for cross-time dependencies and structural relationships across patient records.

Interpretable Deep Neural Networks via Penalization

Jingmao Li, Yale University

Abstract. In the analysis of genomic, imaging, and other high-dimensional complex data, deep neural networks (DNNs) have been widely adopted for their strong representational and predictive capabilities. However, the high dimensionality of inputs, together with noise, poses substantial challenges for model estimation and interpretability. In this paper, we develop a series of DNN-based semiparametric models for high-dimensional settings, in which penalized estimation is employed to facilitate effective variable selection. The proposed methods provides interpretable model structures, principled variable selection procedures, and valid statistical inference. Theoretically, we establish several desirable properties, including estimation consistency, variable selection consistency, and asymptotic normality — results that remain largely underdeveloped in this context. Extensive simulation studies and multiple real data applications demonstrate the strong empirical performance of the proposed methods.

Data Science and the Law (IS-48)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 306

Chair: Neil A. Spencer

Proposer: Neil A. Spencer, University of Connecticut

Presenters: Simon Socolow, Alokesh Manna, Chase Wilson, Sachin S. Pandya

Improving and Evaluating Machine Learning Methods for Forensic Shoeprint Matching

Simon McVey Socolow, Williams College

Abstract. We present SoleMate, a machine learning pipeline for forensic shoeprint matching that improves on the accuracy, computational speed, and interpretability of existing methods. Unlike approaches that focus on class-level distinctions (e.g. make or model), our method examines whether two shoeprints originate from the same shoe based on randomly acquired characteristics (RACs), unique patterns that occur from usage. SoleMate extracts 2D point clouds from shoeprint scans using Laplacian edge detection, aligns the prints using iterative closest point (ICP), and computes a suite of interpretable similarity metrics. These metrics are used to train a random forest classifier that produces probabilistic estimates of whether two prints originated from the same outsole. To assess generalizability, we train and test models on several scenarios: partial prints, prints with varying levels of blurriness, prints with different amounts of wear, and prints from different shoe models. We find that scenario-specific models perform well within their training context but fail to generalize, while our full SoleMate model maintains high accuracy across scenarios. SoleMate achieves higher classification accuracy than recent methods in the literature while maintaining interpretability and being less computationally intensive. To make our work available for other researchers in the field, we release our full codebase, pretrained models, and an interactive web application.

Scalable spatial point process models for large-scale forensic footwear databases

Alokesh Manna, University of Connecticut

Abstract. Shoe print evidence recovered from crime scenes plays a key role in forensic investigations. By examining shoe prints, investigators can determine details of the footwear worn by suspects. However, establishing that a suspect's shoes match the make and model of a crime scene print may not be sufficient. Typically, thousands of shoes of the same size, make, and model are manufactured, any of which could be responsible for the print. Accordingly, a popular approach used by investigators is to examine the print for signs of "accidentals" i.e., cuts, scrapes, and other features that accumulate on shoe soles after purchase due to wear. While some patterns of accidentals are common on certain types of shoes, others are highly distinctive, potentially distinguishing the suspect's shoe from all others. Quantifying the rarity of a pattern is thus essential to accurately measuring the strength of forensic evidence. In this study, we address this task by developing a hierarchical Bayesian model. Our improvement over existing methods primarily stems from two advancements. First, we frame our approach in terms of a latent Gaussian model, thus enabling inference to be efficiently scaled to large collections of annotated shoe prints via integrated nested Laplace approximations. Secondly, we incorporate spatially varying coefficients to model the relationship between shoes' tread patterns and accidental locations. We demonstrate these improvements through superior performance on held-out data, which enhances accuracy and reliability in forensic shoe print analysis.

Seeing Connecticut Through Data with CTData

Chase Wilson, CTData Collaborative

Abstract. An overview of Connecticut Data Collaborative (CTData) including examples of how administrative data from the courts, law enforcement, and legal aid has been analyzed and shared to promote informed policy making in the state. Examples will include what we can learn about Connecticut by analyzing data, and how insight from data can drive important policy in the state. CTData makes public data accessible, builds the data capacity of nonprofits, state agencies, and the public to help make data informed decisions and work toward equity.

Pretrial: A Case Study in Wrangling and Communicating Criminal Justice Data for Policymaking

Sachin S. Pandya, University of Connecticut School of Law

Abstract. When working with court and agency administrative data, data scientists face at least three durable practical challenges: (1) vague or unspecified estimands; (2) the gaps between the administrative data they have and the socio-legal processes that generated that data; and (3) audiences with limited statistical and data visualization literacy. To illustrate these challenges, we recount challenges and

choices in analyzing, for a forthcoming report to the Connecticut Sentencing Commission, criminal court and administrative data to describe the release and detention of the pretrial population (people with pending criminal cases) in Connecticut since 2018.

From Data to Insight in Biomedical and Public Health Research (IS-55)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 111

Chair: Kun Chen

Proposer: Kun Chen, University of Connecticut

Presenters: Chia-Ling Kuo, Youngji Jo, Dongyuan Song, Kun Chen

Bridging Data and Decisions: Statistical Learning from Real-World COVID-19 Surveillance

Youngji Jo, UConn Health

Abstract. Over the past few years, I have collaborated on several COVID-19, related projects with the Korea Disease Control and Prevention Agency and the UK Health Security Agency. In this session, I will highlight two projects. First, in a variant forecasting study, we analyzed over nine million SARS-CoV-2 genomic sequences from GISAID across 15 countries to develop a machine learning, based early warning framework that predicts variant peak prevalence and duration using only 2, weeks of early trajectory data. A key methodological challenge was extracting reliable signals from sparse, noisy, and heterogeneous early genomic data across countries; this was addressed through standardized surveillance thresholds, noise-filtering protocols, and an ensemble Super Learner approach that integrates interpretable epidemiologic features with time-series descriptors. Second, in a nationwide seroprevalence study in South Korea (K-SEROSMART), we integrated serological data (anti-S and anti-N antibodies) with national infection and vaccination surveillance and longitudinal cohort data to estimate infection risk across hybrid, vaccine-only, and infection-induced immunity groups. The primary challenge was substantial under-ascertainment of infections and uncertainty in infection timing; we addressed this by incorporating serology-inferred infections, applying predictive mean matching for imputation, and using time-varying Cox models to capture dynamic immunity effects. Together, these studies illustrate how integrating heterogeneous real-world data sources with tailored statistical methods enables more accurate and policy-relevant inference,Äîfrom early variant detection to characterizing waning population immunity.

Flexible and scalable inference of spatially varying correlation in spatial transcriptomics with spCorr

Dongyuan Song, Department of Genetics and Genome Sciences, University of Connecticut Health Center

Abstract. Spatial transcriptomics has transformed our ability to explore gene expression within its tissue context, enabling us to dissect subtle yet biologically significant variations in situ. While numerous computational methods have been proposed for detecting Spatially Varying Genes (SVGs) expression by modeling each gene separately, much less effort has been devoted to understanding how correlations between genes change across space. Such Spatially Varying Correlations (SVCs) are critical for understanding biological processes such as gene regulatory mechanisms shaped by local tissue environments, yet existing tools remain limited for this task. To address this gap, we present spCorr, a flexible and scalable regression framework for studying SVCs. spCorr provides interpretable, spot-level estimates of gene correlation and detects gene pairs whose correlations vary across locations or between tissue domains. Through extensive simulations and real-data analyses, we show that spCorr achieves high detection power, reliably controls the False Discovery Rate (FDR), and is computationally efficient. Importantly, spCorr reveals biologically meaningful correlation patterns that highlight fine-scale tissue structures, gene module functions, and region-specific interactions, offering new opportunities to study coordinated gene regulation in spatial transcriptomics.

Modern Statistical Learning, Predictive Inference, and Uncertainty Quantification (IS-76)

Time & Location: May 28, 02:00 PM - 03:40 PM | Room 108

Chair: Buxin Su

Proposer: Xingche Guo, University of Connecticut

Presenters: Buxin Su, Anqi Zhao, Yuying Lu, Sicheng Liu

On self-training of summary data with genetic applications

Buxin Su, University of Pennsylvania

Abstract. Prediction model training is often hindered by limited access to individual-level data due to privacy and logistical challenges, particularly in biomedical research. Resampling-based self-training presents a promising approach for building prediction models using only summary-level data. Although this idea has been increasingly implemented in several application fields, its general behaviors remain unexplored. Here we leverage a random matrix theory framework to establish the statistical properties of self-training algorithms for high-dimensional sparsity-free summary data. Notably, we demonstrate that resampling-based self-training can achieve the same asymptotic predictive accuracy as conventional training methods requiring individual-level datasets. These results suggest that self-training with only summary data incurs no additional cost in prediction accuracy while offering practical convenience. Furthermore, we show that the self-training framework maintains this no-cost advantage when combining multiple methods or jointly training on data from different distributions. We validate our findings through simulations and data analyses in the UK Biobank. Our study highlights the potential of resampling-based self-training to advance genetic risk prediction and other fields that make summary data publicly available.

Rethinking Conformal Prediction for Binary Classification

Anqi Zhao, North Carolina State University

Abstract. In binary classification, standard Conformal prediction (CP) often collapses to the uninformative set$\emptyset$or$\{0,1\}$. We identify a structural cause: for any nonconformity score monotone in$\hat{p}(y|x)$, a nontrivial fraction of test points can receive two-label prediction sets. We also show that pointwise level shrinkage$\alpha(x)$under the standard split CP formulation may not achieve conditional validity, yielding a second impossibility result. Motivated by these limits, we propose GLoSaM, a groupwise CP method with data-driven grouping and adaptive calibration that tightens prediction sets while retaining finite-sample distribution-free guarantees. Across synthetic benchmarks and binary LLM and vision classification settings, GLoSaM achieves valid groupwise coverage, substantially higher singleton rates, and robustness to score choice, outperforming e-value and multi-level conformal baselines.

ReFIT: Federated Transfer Learning for Sequential Prediction and Uncertainty Quantification Using Streaming EHR Data

Yuying Lu, Columbia University Mailman School of Public Health

Abstract. Modern biomedical data are increasingly collected across multiple institutions and time periods, creating opportunities for improved statistical inference through knowledge transfer, but also posing challenges for privacy, scalability, and distributional heterogeneity. We propose a Renewable Federated Incremental Transfer framework, termed ReFIT, for sequentially integrating information from streaming source datasets to improve model estimation and prediction in a target population. ReFIT builds upon a density ratio model to account for covariate shift between the source and target populations and employs a renewable updating strategy that allows model parameters to be incrementally refined as new source data become available, using only summary-level information from prior sources. This framework ensures privacy preservation and computational efficiency while adapting to evolving data environments. Beyond improving predictive performance, ReFIT also quantifies predictive uncertainty within a conformal prediction framework, yielding valid prediction intervals that adapt as new information accumulates. Extensive simulation studies demonstrate that ReFIT achieves higher predictive accuracy and better uncertainty quantification than models trained on target or source data alone. The method remains robust even under nonlinear model misspecification and varying degrees of source–target shift. Moreover, as ReFIT incrementally integrates additional source data, the conformal prediction intervals become progressively narrower without sacrificing coverage, evidencing improved statistical efficiency with growing information. In an electronic health record application for breast cancer prediction, ReFIT substantially improves prediction for the Hispanic population by sequentially leveraging information from non-Hispanic White patients collected over multiple time periods. These results highlight the potential of ReFIT as a general and practical framework for privacy-preserving, adaptive, and scalable learning from distributed and periodically updated biomedical data.

Anti-Regularization for Prediction Aggregation

Sicheng Liu, Brown University

Abstract. We study regression settings in which multiple dependent measurements are available for a single target at deployment, and the final prediction is formed by aggregating individual predictions. Motivated by this train/test asymmetry, we introduce the anti-regularized objective

$$M_\alpha(g) = \mathbb{E}[(Y-\overline{g}(Y))^2] + \alpha\,\mathbb{E}[(g(X)-\overline{g}(Y))^2], \quad \overline{g}(Y) = \mathbb{E}[g(X)\mid Y],$$
$$= \mathbb{E}[(Y-g(X))^2] - (1-\alpha)\mathbb{E}[(g(X)-\overline{g}(Y))^2],$$

which reweights the predictive bias–variance decomposition through the parameterα. This loss is naturally induced by deployment with repeated measurements: when the$m$measurements are conditionally i.i.d.~given the target, the mean squared prediction error of aggregation over them equals M_α with some$\alpha \in (0,1)$determined by the aggregation rule. Thus,$M_\alpha$with$0<\alpha<1$provides a population loss for training predictors intended for deployment-time prediction aggregation and can be viewed as an anti-regularized form of mean-squared error. In this work we develop a Hilbert-space analysis of the resulting minimization problem ofM_α, provide conditions for existence/uniqueness of the minimizer, and characterize the minimizer through a first-order optimality condition in the form of an operator equation. Building on this framework, we propose several estimators of$M_\alpha$and training methods for common machine learning models that explicitly align predictors with downstream aggregation. Numerical examples under both the conditional i.i.d.~repeated-measurement model and misspecified repeated-measurement models show that these methods can improve deployment-time error relative to conventional mean-squared-error training.

Parallel Session 4 | 08:45 AM - 10:25 AM, May 29

Advanced Statistical Approaches to Borrow Strength from Real World and Related Subgroups in Clinical Trials (IS-9)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 206

Chair: Ming-hui Chen

Proposer: Kentaro Takeda, Astellas

Presenters: Pang-Yu Liu, Cheng Huang, Kentaro Takeda, Wei Wei

A comparison of information borrowing methods in basket trials

Pang-Yu Liu, UCONN

Abstract. Basket trials are increasingly used in early-phase oncology drug development to evaluate biomarker-targeted therapies in multiple cancer cohorts. Bayesian approaches are particularly popular in this setting, as their hierarchical structures can be adapted to allow information borrowing. This paper reviews recent advances in Bayesian basket trial methodologies. Specifically, we systematically compare a spectrum of established statistical designs—including independent analysis (no borrowing), Bayesian Hierarchical Models (BHM), EXNEX, Local Power Prior (LPP), Bayesian Model Averaging (BMA), and Mixture of Finite Mixtures (MFM) with proposed prior. Within this scope, we summarize the strengths, limitations, underlying differences, and theoretical connections of these approaches. In addition, extensive simulation studies are conducted to examine the empirical performance of all of the methods considered. Finally, real data from the Vemurafenib trial are further analyzed to demonstrate the practical utility of these methods.

Information Borrowing in Basket Trials Using the Bayesian Model Average Method

Cheng Huang, Vir Biotechnology

Abstract. Basket trials have become increasingly accepted and widely adopted study designs in disease areas, particularly in oncology and rare diseases, due to their advantages in operational efficiency and ethics. Bayesian methods are commonly applied in these trials, as Bayesian hierarchical structures naturally facilitate information borrowing across baskets. Among these approaches, a key aspect is whether to borrow information globally or locally, which leads to different borrowing strategies, such as hierarchical models (global borrowing) and mixture models (local borrowing). Bayesian model averaging provides a flexible framework that incorporates multiple candidate models and adjusts their relative weights based on the observed data. In this work, we present simulation analyses to compare Bayesian methods, including Bayesian model averaging, under scenarios with continuous endpoints and multiple stages. The Family-wise

error rate and power are the primary criteria used to evaluate and compare the methods.

A randomized basket trial design for dose optimization based on Bayesian model averaging using spike-and-slab priors

Kentaro Takeda, Astellas

Abstract. The FDA initiated Project Optimus and issued guidance for dose optimization, recommending randomized parallel dose-response cohorts to generate additional data at promising dose levels and implies that different dosages may be needed for different indications. In addition to dose optimization, with recent advancements in precision medicine and cancer biology, the development of cancer treatments has shifted toward the search for agents targeted to specific molecular profiles that may appear in more than one type of cancer. The basket trials are clinical trial designs that enable the simultaneous assessment of a new treatment in multiple indications. Concerning the FDA's dose optimization perspective and the recent trend of basket trials in early-phase clinical trials, this paper proposes a dose-ranging basket trial design based on a Bayesian model-averaging approach considering efficacy and toxicity outcomes, where indications and dose levels define baskets. A key benefit of the proposed approach is that it explicitly accounts for the possible heterogeneity of response rates among baskets. Our simulation study shows that the proposed approach outperforms other methods, offering higher statistical power, better control of Type I error rates, precise optimal dose selection, and sample size savings in various scenarios with heterogeneous treatment effects between baskets.

Learning from Literature: Integrating LLMs and Bayesian Hierarchical Modeling for Oncology Trial Design

Wei Wei, Yale University

Abstract. Designing modern oncology trials requires synthesizing evidence from prior studies to inform hypothesis generation and sample size determination. Trial designs based on incomplete or imprecise summaries can lead to misspecified hypotheses and underpowered studies, resulting in false positive or negative conclusions. To address this challenge, we developed LEAD-ONC (Literature to Evidence for Analytics and Design in Oncology), an AI-assisted framework that transforms published clinical trial reports into quantitative, design-relevant evidence. Given expert-curated trial publications that meet prespecified eligibility criteria, LEAD-ONC uses large language models to extract baseline characteristics and reconstruct individual patient data from Kaplan-Meier curves, followed by Bayesian hierarchical modeling to generate predictive survival distributions for a prespecified target trial population. We demonstrate the framework using five phase III trials in first-line non-small-cell lung cancer evaluating PD-1 or PD-L1 inhibitors with or without CTLA-4 blockade. Clustering based on baseline characteristics identified three clinically interpretable populations defined by histology. For a prospective randomized trial in the mixed-histology population comparing mono versus dual immune checkpoint inhibition, LEAD-ONC projected a modest median overall survival difference of 2.8 months (95 percent credible interval -2.0 to 7.6) and an estimated probability of at least a 3-month benefit of approximately 0.45. As LEAD-ONC remains under active development, these results are intended as preliminary demonstrations of the frameworks potential to support evidence-driven oncology trial design rather than definitive clinical conclusions.

Invited Papers from the New England Journal of Statistics in Data Science (NEJSDS) - Recent developments on real-world data science, variable selection and regression machine learning (IS-13)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 101

Chair: HaiYing Wang

Proposer: Colin O. Wu, National Institutes of Health

Presenters: Jun Yan, Eric Odoom, Sam Hawke, Colin Wu

Data Jamboree and Beyond: From Programming to Reproducible Research

Jun Yan, University of Connecticut

Abstract. The growing emphasis on computing in statistics and data science education has increased the need for training that integrates statistical reasoning with modern computational workflows. This presentation discusses the Data Jamboree, a live educational event that combines hands-on instruction with real-world data science projects using open data. Participants ranging from high school students

to advanced users followed workshop leaders in using open-source tools such as R, Python, and Julia for data cleaning, visualization, modeling, and predictive analysis. The presentation further extends beyond the original paper by introducing modern tools for reproducible and collaborative research, including Quarto, Git, and GitHub. These tools are presented as foundational components of contemporary data science workflows that support version control, reproducibility, scientific communication, and collaborative development. The presentation concludes with recommendations for organizing educational activities that prepare students for research-oriented and team-based data science

Consistent and Scalable Variable Selection with Robust Link Functions

Eric Odoom, University of Cincinnati

Abstract. This study explores the application of the t-link model in high-dimensional variable selection for binary regression. The t-link model provides flexibility in binary modeling and offers robust inference in the presence of outliers, making it a preferable alternative to the commonly used probit and logit links. To address the computational challenges posed by a large number of covariates, the skinny Gibbs algorithm is employed, and the consistency of variable selection under this approximate algorithm is established. These advancements in both computational and theoretical perspectives enhance the practicality and ease of implementing the t-link model. The performance of the t-link model, with a specified degrees of freedom, is compared with the logit and probit links through simulation studies and an application to PCR data. The results demonstrate the robustness and computational efficiency of the proposed method.

Contrastive Inverse Regression for Dimension Reduction

Sam Hawke, Skidmore College

Abstract. Contrastive dimension reduction (CDR) methods aim to extract signal that is unique to or enriched in a target group relative to a background group, a setting that arises naturally in case-control studies across genomics, imaging, and other high-dimensional data domains where standard methods such as PCA may fail to isolate the signal of interest. In this talk, I will give a systematic overview of the CDR landscape, introducing a taxonomy of existing methods organized by their assumptions, objectives, and mathematical formulations. I will then present contrastive inverse regression (CIR), a supervised CDR method that preserves the functional relationship between dimension-reduced covariates and a response variable. CIR poses an optimization problem on the Stiefel manifold and is shown to converge to a local optimum via a gradient descent-based algorithm. I will demonstrate its performance on benchmark datasets from single-cell genomics, including the BMMC and COVID-19 cell states datasets, and discuss open questions and future directions in this emerging field.

Recent Breakthroughs in Nonparametric and Resampling-Based Inference Under Dependence (IS-14)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 205

Chair: Haihan Yu

Proposer: Haihan Yu, University of Rhode Island

Presenters: Yingchao Zhou, Anderson Bussing, Souvick Bera, Haihan Yu

A resampling-based method for accurate Whittle intervals with nonlinear time series

Yingchao Zhou, Iowa State University

Abstract. Whittle estimation is an important problem for time series, which involves fitting covariance models to data in order to understand dependence structure. However, uncertainty quantification for Whittle estimators has remained challenging and existing interval estimators often suffer poor coverage, which owes to issues in bias and complex variances. This talk describes a new approach for calibrating nonparametric confidence intervals for Whittle parameters based on a combination of empirical likelihood and bootstrap. That is, the method allows spectral density (i.e., covariance) models to misspecified and applies to general, possibly non-linear, time series. Numerical evidence illustrates good accuracy of the proposed resampling-based method.

A Risk Sharing Rule for Peer-to-peer (P2P) Insurance for Extreme Weather Losses via Conditional Expectation

Anderson Bussing, University of South Carolina

Abstract. We consider risk-sharing in peer-to-peer (P2P) hail insurance, where participants pool their losses and contribute ex-post according to a sharing rule. The rule we adopt is the conditional mean

risk-sharing rule of Denuit and Dhaene (2012), under which each participant contributes the conditional expectation of their loss given the pool's total. This rule ensures actuarial fairness— on average, the operation does not transfer contributions from one participant to the others— along with several other desirable properties. Computing the rule requires a joint model for the individual losses and a method for evaluating the resulting conditional expectations under dependence. In hail insurance, losses are sparse, spatially correlated within storm events, and heterogeneous across property-level features. We propose a joint frequency-severity model with a shared latent Gaussian process at the storm-date level, and compute the conditional expectations by extending the procedure of Denuit and Robert (2021). We demonstrate the framework on a Nebraska hail insurance dataset.

Frequency Domain Resampling for Gridded Spatial Data

Souvick Bera, Colorado School of Mines

Abstract. In frequency domain analysis for spatial data, spectral averages based on the periodogram often play an important role in understanding spatial covariance structure, but also have complicated sampling distributions due to complex variances from aggregated periodograms. In order to non-parametrically approximate these sampling distributions for purposes of inference, resampling can be useful, but previous developments in spatial bootstrap have faced challenges in the scope of their validity, specifically due to issues in capturing the complex variances of spatial spectral averages. As a consequence, existing frequency domain bootstraps for spatial data are highly restricted in application to only special processes (e.g. Gaussian) or certain spatial statistics. To address this limitation and to approximate a wide range of spatial spectral averages, we propose a practical hybrid-resampling approach that combines two different resampling techniques in the forms of spatial subsampling and spatial bootstrap. Subsampling helps to capture the variance of spectral averages while bootstrap captures the distributional shape. The hybrid resampling procedure can then accurately quantify uncertainty in spectral inference under mild spatial assumptions. Moreover, compared to the more studied time series setting, this work fills a gap in the theory of subsampling/bootstrap for spatial data regarding spectral average statistics.

Nonparametric Inference under Gaussian Subordinated Process

Haihan Yu, University of Rhode Island

Abstract. Long-range dependent time series$\{X_t\}$can often be embedded in the Gaussian subordinated process, regarding which as an unknown transformation$G(\cdot)$of a long-memory Gaussian process$\{Z_t\}$(i.e.,$X_t = G(Z_t)$). Despite substantial methodological developments for long-memory time series over the past 50 years, inference for such possibly nonlinear processes remains largely an open problem. This difficulty stems from the inaccessibility of key parameters, the Hermite rank$m$of$G$and the latent memory exponent$\alpha$of$\{Z_t\}$, as well as the non-Gaussian limiting distributions that commonly arise under long memory. We introduce the first formal method for jointly estimating$(m, \alpha)$and establish its consistency under mild conditions. Building on these estimators, we propose a broadly applicable bootstrap-based inference framework, which we exemplify for the mean, approximately linear statistics, and the empirical distribution. The performance of the proposed methods is assessed via simulation studies and their practical utility is illustrated through applications to the Dow Jones Index.

Recent Advances in Design of Experiments (IS-23)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 305

Chair: Chenlu Shi

Proposer: Chenlu Shi and Haiying Wang, New Jersey Institute of Technology, University of Connecticut

Presenters: Chenlu Shi, Frederick Phoa, Chi-Kuang Yeh, Chao-Hui Huang

Axis-Neighbor Geometry and Length-Scale Identifiability in Gaussian Process Models

Chenlu Shi, New Jersey Institute of Technology

Abstract. In computer experiments, experimental designs for Gaussian process surrogates are often chosen to achieve global space-filling coverage for prediction. However, the same simulator runs must also support estimation of covariance parameters, and designs that are effective for prediction alone may yield weak or unstable length-scale inference. This issue is especially acute for anisotropic GP models, where

each coordinate has its own length-scale and informative comparisons are inherently directional. We develop a geometry-driven framework that links experimental design directly to length-scale identifiability in anisotropic GP models with Gaussian correlation. For general designs, we derive lower bounds for the diagonal Fisher information that depend explicitly on local axis-neighbor geometry through the minimum coordinate-isolating spacing and the multiplicity of closest such pairs. These results show that, in the small-length-scale regime, stable inference is governed by local directional structure rather than by global spread alone. Motivated by this theory, we propose a structured design that is a strong orthogonal array of strength 2+, and hence combines space-filling properties with a rich collection of short coordinate-isolating pairs. We further derive explicit Fisher-information bounds and sharp small-length-scale asymptotics for the proposed design, and introduce a structured level-expansion that refines global resolution while preserving the local geometry responsible for length-scale identifiability. Numerical studies indicate that the proposed designs substantially improve the stability of length-scale estimation while maintaining competitive predictive performance across a range of length-scale regimes.

Optimal Semi-Supervised Subsampling for Softmax Regression

Frederick Kin Hing Phoa, Institute of Statistical Science, Academia Sinica

Abstract. Missing label information presents a significant challenge for optimal subsampling methods, which typically rely on complete response data to compute sampling probabilities. In this study, we propose a semi-supervised A-/L-optimal subsampling framework for softmax regression that effectively addresses this issue. We derive the optimal subsampling probabilities under the baseline constraint and highlight their role in balancing categorical responses. In addition, we explore constraint-invariant subsampling by minimizing the asymptotic mean squared prediction error (MSPE), enabling the construction of subsampling probabilities for each observation, which is robust to model constraint choices. Our theoretical findings are supported by simulations and real-data applications, demonstrating improvements in both prediction accuracy and computational efficiency.

Single and multi-objective optimal designs for group testing experiments with a focus on screening for an infectious disease

Chi-Kuang Yeh, Georgia State University

Abstract. Group testing, or pooled-sample testing, is widely used in large-scale screening and resource constrained studies, yet principled design methodology for precise parameter estimation remains limited. This talk presents an optimal design framework for group testing that targets efficient estimation of key model parameters while accounting for cost constraints. Beyond classical criteria such as D-, Ds-, and A-optimality, a central novelty is the introduction of maximin design principles, including potentially non-differentiable criteria, into group testing procedures. These non-differentiable criteria have not previously been explored in this context and yield designs with strong worst-case guarantees and improved robustness. The framework accommodates both large-sample settings through optimal approximate designs and small-sample studies through exact optimal designs, enabling systematic assessment of robustness to changes in criteria, statistical models, and cost structures. We demonstrate the practical impact of this approach through an application to Chlamydia screening using imperfect assays under budget constraints, and show that precise parameter estimation via optimal design is a foundational step that directly enables efficient and reliable sequential group-testing procedures.

A Multi Objective Optimal Design Framework for Two-Sample Sparse Functional Data

Chao-hui Huang, Institute of Statistical Science, Academia Sinica, Taiwan

Abstract. We study optimal design for two-sample functional data, where the goals of accurately recovering individual trajectories and conducting powerful tests for mean function differences often conflict. We introduce two design criteria—one maximizing the noncentrality parameter of a projection-based two-sample test, and the other minimizing the mean integrated squared error for predicting curve differences. To jointly optimize these objectives, we propose MOSIB, a multi-objective swarm intelligence–based algorithm that efficiently estimates the Pareto front. A theoretical result shows that power-optimal designs can be searched within single-support designs, substantially reducing computation. Simulation studies demonstrate that MOSIB consistently identifies high-quality designs, recovers single-objective op-

tima, and produces superior Pareto fronts compared with random search. An application to earthquake signal data further highlights that optimized designs can detect subtle between-group differences while maintaining high trajectory estimation accuracy. The proposed framework provides a flexible and computationally efficient approach to multi-objective design in sparse functional data analysis.

Advances in Win Statistics for Randomized Trials (IS-26)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 301

Chair: Guangyu Tong

Proposer: Guangyu Tong, Yale University

Presenters: Lu Mao, Xi Fang, Kenneth Lee

Regularized win ratio regression for variable selection and risk prediction, with an application to a cardiovascular trial

Lu Mao, University of Wisconsin-Madison

Abstract. The win ratio has been widely used in the analysis of hierarchical composite endpoints, which prioritize critical outcomes such as mortality over nonfatal, secondary events. Although a regression framework exists to incorporate covariates, it is limited to low-dimensional datasets and may struggle with numerous predictors. This gap necessitates a robust variable selection method tailored to the win ratio framework. We propose an elastic net-type regularization approach for win ratio regression, extending the proportional win-fractions (PW) model in low-dimensional settings. The method addresses key challenges, including adapting pairwise comparisons to penalized regression, optimizing model selection through subject-level cross-validation, and defining performance metrics via a generalized concordance index. The procedures are implemented in the wrnet R-package, publicly available at https://lmaowisc.github.io/wrnet/. Simulation studies demonstrate that wrnet outperforms traditional (regularized) Cox regression for time-to-first-event analysis, particularly in scenarios with differing covariate effects on mortality and nonfatal events. When applied to data from the HF-ACTION trial, the method identified prognostic variables and achieved superior predictive accuracy compared to regularized Cox models, as measured by overall and component-specific concordance indices.

Generalized win fraction regression for composite survival endpoints

Xi Fang, Medical College of Wisconsin

Abstract. We propose a regression framework for studying pairwise relative effect measures based on composite time-to-event outcomes subject to right censoring. Our model expands the probabilistic index model from univariate uncensored outcomes to composite survival outcomes by directly modeling the conditional win fraction through a prespecified link function. To accommodate right censoring, we construct inverse probability of censoring weighted estimating equations for consistent estimation of regression coefficients. The proposed regression framework accommodates multiple association effect measure scales. Under an identity link, regression parameters characterize the covariate associations on the natural scale of the win fraction scale over follow-up, whereas under a logit link without ties, our model recovers the proportional win fraction regression and reflects associations on the win ratio scale. Large-sample properties of the regression coefficient estimators are established, and a consistent sandwich variance estimator accounting for the uncertainty in estimating the censoring weights is provided. We carry out extensive simulation studies to examine the finite-sample performance of the generalized win fraction regression approach, and further illustrate our this new method by re-analysis of the HF-ACTION clinical trial.

Who's Winning? Clarifying Estimands Based on Win Statistics in Cluster Randomized Trials

Kenneth Lee, University of Pennsylvania

Abstract. Treatment effect estimands based on win statistics, including the win ratio, win odds, and win difference are increasingly popular targets for summarizing endpoints in clinical trials. Such win estimands offer an intuitive approach for prioritizing outcomes by clinical importance. The implementation and interpretation of win estimands is complicated in cluster randomized trials (CRTs), where researchers can target fundamentally different estimands on the individual-level or cluster-level. We numerically demonstrate that individual-pair and cluster-pair win estimands can substantially differ when cluster size is informative: where outcomes and/or treatment effects depend

on cluster size. With such informative cluster sizes, individual-pair and cluster-pair win estimands can even yield opposite conclusions regarding treatment benefit. We describe consistent estimators for individual-pair and cluster-pair win estimands and propose a leave-one-cluster-out jackknife variance estimator for inference. Despite being consistent, our simulations highlight that some caution is needed when implementing individual-pair win estimators due to finite-sample bias. In contrast, cluster-pair win estimators are unbiased for their respective targets. Altogether, careful specification of the target estimand is essential when applying win estimators in CRTs. Failure to clearly define whether individual-pair or cluster-pair win estimands are of primary interest may result in answering a dramatically different question than intended.

Causal Inference and Adaptive Design under Clustering and Interference (IS-27)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 111

Chair: Guangyu Tong

Proposer: Guangyu Tong, Yale School of Medicine

Presenters: Yuki Ohnishi, Jingyu Cui, Fang Fei, Oliver Hines

Identification and estimation of causal mechanisms in cluster-randomized trials with post-treatment confounding using Bayesian nonparametrics

Yuki Ohnishi, Yale School of Public Health

Abstract. Causal mediation analysis in cluster-randomized trials (CRTs) is essential for explaining how cluster-level interventions affect individual outcomes, yet it is complicated by interference, post-treatment confounding, and hierarchical covariate adjustment. We develop a Bayesian nonparametric framework that simultaneously accommodates interference and a post-treatment confounder that precedes the mediator. Identification is achieved through a multivariate Gaussian copula that replaces crossworld independence with a single dependence parameter, yielding a built-in sensitivity analysis to residual post-treatment confounding. For estimation, we introduce a nested common atoms enriched Dirichlet process (CA-EDP) prior that integrates the Common Atoms Model (CAM) to share information across clusters while capturing between- and within-cluster heterogeneity, and an Enriched Dirichlet Process (EDP) structure delivering robust covariate adjustment without impacting the outcome model. We provide formal theoretical support for our prior by deriving the model's key distributional properties, including its partially exchangeable partition structure, and by establishing convergence guarantees for the practical truncationbased posterior inference strategy. We demonstrate the performance of the proposed methods in simulations and provide further illustration through a reanalysis of a completed CRT.

Analysis of Learn-As-you-GO (LAGO) in stepped wedge design with center random effects

Jingyu Cui, Yale School of Public Health

Abstract. Implementation science studies often fail to achieve their intended outcomes, raising an important question: how can limited resources be allocated to maximize effectiveness while minimizing costs? In this article, we introduce the Learn-As-you-GO (LAGO) adaptive design within the framework of the stepped wedge design (SWD). Under LAGO, interventions in later stages are updated using accumulating data from earlier stages to improve cost-efficiency. This adaptive structure induces complex dependencies across stages, centers, and patients, posing challenges for classical statistical methods that typically rely on independent and identically distributed observations. We show that estimators obtained from the generalized estimating equation approach retain their classical asymptotic properties when LAGO is implemented in an SWD. Simulation studies demonstrate that LAGO can achieve a better balance between desired outcomes and study costs than fixed designs. An analysis of the BetterBirth study provides a real-world illustration of how interventions may be adaptively updated during a trial. Overall, these results highlight the potential of LAGO to advance implementation science.

Adaptive Experimental Design for Efficient Causal Estimators under Neighborhood and Temporal Interference

Fei Fang, Yale University

Abstract. Network interference poses fundamental challenges for experimental design, as incompatibilities among nearby units in simultaneously attain-

ing the exposure conditions defining the target estimand can result in low estimation efficiency. We develop an adaptive design for estimating causal effects under neighborhood interference, with a focus on improving estimator efficiency. We first consider experiments conducted on multiple networks, where the network structure, potential outcomes, and covariates are sampled from a common distribution over time. In this regime, we build on the conflict graph design of Kandiros et al. (2025), which reduces conflicts in realizing target exposure conditions by assigning treatment according to an importance ordering of neighboring units. In the proposed adaptive conflict graph design, exposure sampling probabilities and importance orderings are updated based on observed history and are chosen to minimize estimator variance. We then study adaptive design on a single fixed network observed repeatedly over time, where temporal carryover as well as dependence arise. To address this setting, we propose a block-adaptive design that applies the adaptive conflict graph design at the beginning of each block, with estimation leveraging observations from the carryover period. We jointly minimize estimator variance over the block length, exposure sampling probabilities, and importance orderings in an adaptive manner. Simulation studies and synthetic data applications demonstrate efficiency gains relative to the best non-adaptive benchmarks.

Where best to intervene?

Oliver Hines, Columbia University

Abstract. Mediation analysis is often used to find avenues for future interventions that may improve the efficacy of a exposure or mitigate harmful effects or disparity. That is, mediation analysis is thought to help address the question "Where best to intervene?' '. In this talk, we formalize this notion. This naturally leads to causal estimands that inform about the effects of future (generally stochastic) interventions on unobserved variables along causal pathways. Some of these interventions correspond to randomized interventional analogs of natural direct and indirect effects and path-specific effects, and some correspond to so-called separable effects. Unlike "natural' ' mediation effects, our effects remain manipulable in the sense that they follow the target trial principle: there is a hypothetical randomized trial in which they could be estimated. Moreover, identification does not rely on cross-world counterfactual assumptions, which are often criticized for their lack of interpretability.

Recent advances in statistical methods for analyzing high dimensional biomedical data (IS-35)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 201

Chair: Qi Zhang

Proposer: Qi Zhang, University of New Hampshire

Presenters: Qian Zhao, Jiang Gui, Zhou Lan, Yunlong Feng

Controlled variable selection with a biased sample using tilted knockoffs

Qian Zhao, University of Massachusetts Amherst

Abstract. Researchers in biomedical studies often work with biased samples that are not selected uniformly at random from the population of interest. One example is a case-control study, where cases are over-sampled to study risk factors of rare diseases. While these designs are motivated by specific scientific questions, it is often of interest to use them to pursue secondary lines of investigations. In these cases, the biased sample can lead to spurious association between an exposure and an outcome when both of them affect the case-control status. This phenomenon is known in the causal inference literature as collider bias. While tests of independence under biased sampling are available, these methods typically do not apply when the number of variables is large. In this work, we are interested in using the knockoff framework to select important variables among very many with replicability guarantees. We show that the standard model-X knockoffs fail to control FDR in the presence of biased sampling. We show that by tilting the population distribution with the selection probability and constructing knockoff variables according to this tilted distribution, the knockoff filter would control the FDR. We apply the tilted knockoff method to identify genetic underpinning of endophenotypes in a case-control study.

HiST: A Hierarchical Sparse Transformer for Cross-Modal Spatial Transcriptomics Modeling

Jiang Gui, Dartmoouth College

Abstract. Spatial transcriptomics (ST) links gene expression with tissue morphology but remains expensive and low-throughput, motivating surrogates that infer expression from routine histology. Whole-slide H&E-to-ST inference pairs a gigapixel

image with gene measurements at a sparse, irregular set of locations, making multiscale modeling challenging without incurring dense-grid overhead or quadratic token mixing. We propose HiST, a hierarchical sparse transformer that treats measured locations as a lattice-indexed sparse field and builds a dyadic encoder–decoder directly on the active tissue footprint. HiST combines sparse window attention for local geometric correspondence with resolution-changing operators for rapid multiscale context integration. For a fixed window size, the dominant runtime and memory scale with the number of observed locations rather than the dense slide area. To mitigate slide-specific acquisition variation, HiST adds a bottlenecked global conditioning pathway via a *slide calibration token* that summarizes slide-level context and conditions local representations. On a multi-organ benchmark spanning diverse tissues and acquisition sources, HiST improves predictive performance over recent baselines while reducing runtime and peak memory.

Statistical Innovation in Neuroimaging Biomarker Development: The Neurometabolic Network (NMetNet) for a Neurological Disorder

Zhou Lan, Sanofi

Abstract. Functional neurological disorder (FND) in children and adolescents is a complex condition shaped by interacting biological, psychological, and social factors. It presents with a diverse range of neurological symptoms and has increasingly been studied using advanced neuroimaging methods. One such method, magnetic resonance spectroscopy, allows researchers to examine brain chemistry by measuring neurometabolites. While earlier studies primarily focused on the concentrations of these metabolites, this study instead investigates how they relate to each other through conditional dependencies. Specifically, it examines six key neurometabolites—N-acetyl aspartate, creatine, glutathione, choline, myo-inositol, and glutamate—and defines conditional dependence as the shared variability between two metabolites that cannot be explained by the influence of others. To analyze these relationships, the study uses a Bayesian graphical lasso method to estimate conditional dependencies across three brain regions: the anterior default mode network (aDMN), the supplementary motor area (SMA), and the posterior default mode network (pDMN). The resulting framework is referred to as the neurometabolic network (NMetNet), which captures the interconnected structure of metabolite relationships rather than isolated levels. The findings reveal that children and adolescents with FND, compared to healthy controls, show a disruption in conditional dependencies involving creatine and glutathione, particularly between the aDMN and the SMA or pDMN. Glutathione is known as the brain's primary antioxidant, while creatine plays a critical role in maintaining cellular energy balance and also contributes to antioxidant defense. Overall, these results suggest that FND is associated with dysregulation in brain energy metabolism and an increased susceptibility to oxidative stress. By examining neurometabolic networks rather than individual metabolite levels, this approach provides new insights into the neurobiological underpinnings of FND and highlights potential therapeutic targets aimed at restoring energy homeostasis and reducing oxidative stress.

Data privacy: theory and practice (IS-40)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 302

Chair: Nianqiao Phyllis Ju

Proposer: Nianqiao Phyllis Ju, Dartmouth College

Presenters: Carlos Soto, Quanquan Liu, Jonathan Ullman, Daniel Sheldon

Rao Differential Privacy

Carlos Soto, University of Massachusetts Amherst

Abstract. Differential privacy (DP) has recently emerged as a definition of privacy. Summary statistics are sanitized by injecting noise which satisfies DP. DP calibrates noise to be on the order of an individual's contribution. Due to this calibration, a private estimate obscures any individual while preserving the utility of the estimate. Since the original definition, many alternate definitions have been proposed. These alternates have been proposed for various reasons including improvements on composition results, relaxations, and formalizations. Nevertheless, thus far nearly all definitions of privacy have used a divergence of densities as the basis of the definition. In this paper we take an information geometry perspective towards differential privacy. Specifically, rather than define privacy via a divergence, we define privacy, Rao Differential Privacy, via the Rao distance. We show that our proposed definition of privacy shares the interpretation of previous definitions of privacy

while improving on sequential composition. Further, we demonstrate the relationship between Rao DP and many classical differential privacy definitions.

Practical and Accurate Local Edge Differentially Private Graph Algorithms

Quanquan Liu, Yale University

Abstract. The rise of massive networks across diverse domains necessitates sophisticated graph analytics, often involving sensitive data and raising privacy concerns. This paper addresses these challenges using local differential privacy (LDP), which enforces privacy at the individual level, where no third-party entity is trusted, unlike centralized models that assume a trusted curator. We introduce novel LDP algorithms for two fundamental graph statistics: k-core decomposition and triangle counting. Our approach leverages input-dependent private graph properties, specifically the degeneracy and maximum degree of the graph, to improve theoretical utility. Unlike prior methods, our error bounds are determined by the maximum degree rather than the total number of edges, resulting in significantly tighter guarantees. For triangle counting, we improve upon the work of Imola, Murakami, and Chaudhury [USENIX Security 21,22], which bounds error in terms of edge count. Instead, our algorithm achieves bounds based on graph degeneracy by leveraging a private out-degree orientation, a refined variant of Eden et al.'s randomized response technique [ICALP 23], and a novel analysis, yielding stronger guarantees than prior work. Beyond theoretical gains, we are the first to evaluate local DP algorithms in a distributed simulation, unlike prior work tested on a single processor. Experiments on real-world graphs show substantial accuracy gains: our k-core decomposition achieves errors within 3x of exact values, far outperforming the 131x error in the baseline of Dhulipala et al. [FOCS22]. Our triangle counting algorithm reduces multiplicative approximation errors by up to six orders of magnitude, while maintaining competitive runtime.

The Sample Complexity of Membership Inference and Privacy Auditing

Jonathan Ullman, Northeastern University

Abstract. A membership-inference attack gets the output of a learning algorithm, and a target individual, and tries to determine whether this individual is a member of the training data or an independent sample from the same distribution. A successful membership-inference attack typically requires the attacker to have some knowledge about the distribution that the training data was sampled from, and this knowledge is often captured through a set of independent reference samples from that distribution. In this work we study how much information the attacker needs for membership inference by investigating the sample complexity-the minimum number of reference samples required-for a successful attack. We study this question in the fundamental setting of Gaussian mean estimation where the learning algorithm is given n samples from a Gaussian distribution in d dimensions, and tries to estimate its mean up to some error . Our result shows that for membership inference in this setting, many more than n samples can be necessary to carry out any attack that competes with a fully informed attacker. Our result is the first to show that the attacker sometimes needs many more samples than the training algorithm uses to train the model. This result has significant implications for practice, as all attacks used in practice have a restricted form that uses O(n) samples and cannot benefit from (n) samples. Thus, these attacks may be underestimating the possibility of membership inference, and better attacks may be possible when information about the distribution is easy to obtain.

Private Adaptive Covariance Estimation via Gaussian Graphical Models

Daniel Sheldon, University of Massachusetts Amherst

Abstract. Most existing methods for estimating a covariance matrix under differential privacy add noise to each entry of the empirical covariance matrix. We investigate whether it can be more effective to adaptively measure a subset of entries and then reconstruct a full matrix. Under natural assumptions, individual entries can be measured more precisely than the full matrix for the same privacy cost, so this strategy can focus its measurements on more informative entries. In each round, our method selects a poorly approximated entry, measures it using the Gaussian mechanism, and then reconstructs a full covariance matrix. We formulate a maximum-entropy reconstruction problem that generalizes Dempster's covariance selection, develop efficient algorithms to solve it, and show that solutions correspond to Gaussian graphical models. Experiments demonstrate consistent improvements in estimation error compared to the Gaussian mechanism and other baselines,

particularly in high dimensions and stricter privacy settings.

Statistical learning and computation for PDEs and non-linear systems: from uncertainty quantification to deep neural networks for parameter estimation. (IS-42)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 202

Chair: Julio Enrique Castrillon

Proposer: Debarghya Mukherjee, Boston University

Presenters: Debarghya Mukherjee, Akshunna Shaurya Dogra, Jie Xu, Julio Enrique Castrillon

Learning PDE-constrained inverse problems via frequentist and Bayesian debiasing

Debarghya Mukherjee, Boston University

Abstract. We study the problem of estimating unknown parameters in PDE-constrained inverse problems from noisy observations, where the PDE solution is approximated using Physics-Informed Neural Networks (PINNs). While PINNs have demonstrated remarkable empirical success, existing estimators typically inherit the slow nonparametric convergence rate of the neural network solution, resulting in biased and statistically inefficient inference on the finite-dimensional parameters of interest. To address this, we propose a two-step debiased estimation procedure that combines neural network-based nonparametric estimation with an influence function-based bias correction. By eliminating the first-order sensitivity of the estimator to errors in the nuisance function, our procedure yields a$\sqrt{n}$-consistent and asymptotically normal estimator without requiring undersmoothing of the neural network component. We further extend this framework to a Bayesian inference by replacing the original likelihood with a debiased quasi-likelihood, and establish a Bernstein–von Mises theorem showing that the resulting posterior contracts at the$\sqrt{n}$-rate with an asymptotic covariance matching that of the frequentist estimator. As a by-product of our analysis, we establish near-minimax optimal convergence rates for estimating a nonparametric regression function and its derivatives in Sobolev spaces using neural networks. Extensive numerical experiments corroborate our theoretical findings and demonstrate the necessity of the proposed debiasing procedure for valid statistical inference in PDE-constrained inverse problems.

Neural ODE/PDE with Complex Analyticity

Jie Xu, Northeastern University

Abstract. Recently, Neural ODE/PDE methods like PINN, NO, DeepONet, etc. are very popular in approximating, simulating or learning the solutions of PDE/PDEs. In particular, these machine learning methods provide efficient alternatives to classical methods (FEM, etc.) for complex, high-dimensional, and time-dependent nonlinear systems. However, the learning qualities of Neural ODE/PDE methods rely on the scarcity of the data, the spectral bias, nonlinearity, depth and width of the neural networks, etc. and therefore may still be computationally heavy when pursuing desired level of accuracy. In this talk, we show how the combination of holomorphic extensions of (Stochastic) ODE/PDEs and complexified neural network methods partially resolve these issues.

Uncertainty Quantification for stochastic non-linear PDEs and networks

Julio Enrique Castrillon, Boston University

Abstract. Nonlinear stochastic partial differential equations and nonlinear network models arise in a wide range of scientific and engineering applications, including power systems, electrostatics, computational chemistry, and biophysical modeling. In high-dimensional settings, the computation of statistical quantities of interest becomes prohibitively expensive due to the curse of dimensionality, particularly when conventional Monte Carlo methods are used at these scales. This motivates the development of scalable uncertainty quantification (UQ) methodologies capable of exploiting additional mathematical structure in the underlying stochastic systems. In this talk, we present recent advances in the analysis and numerical approximation of high-dimensional nonlinear stochastic systems through the use of complex analytic regularity and sparse approximation techniques. The central theme is that, under suitable stochastic parameterizations, solutions of nonlinear equations admit analytic extensions with respect to the random variables in complex domains. This analytic structure enables the application of sparse-grid and related high-dimensional approximation methods with algebraic to sub-exponential convergence rates, thereby substantially mitigating the curse of

dimensionality. Applications are presented for nonlinear stochastic network models and the nonlinear Poisson–Boltzmann equation (NPBE), including uncertainty quantification under random perturbations of loads, coefficients, and geometries. Using techniques based on the analytic implicit function theorem, domain mapping methods, and contraction mapping arguments, we establish analyticity of the solutions with respect to high-dimensional random parameters and derive corresponding sparse approximation results. Numerical experiments involving power-system models and biomolecular electrostatics demonstrate convergence rates consistent with the theory and dramatic computational improvements over conventional Monte Carlo methods.

Careers in Academia: A Panel Discussion by NESS NextGen (IS-45)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 102

Chair: Elizabeth Upton

Proposer: Dr. Elizabeth Upton, Dr. Gregory Vaughan, Williams College; Bentley University

Panelists: Benjamin Seiler, Jungwun Lee, Mihaela Predescu, Eddie Kim

New Advances in Causal Inference and Data Science (IS-53)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 306

Chair: Ying Zhou

Proposer: Ying Zhou, University of Connecticut

Presenters: Yubai Yuan, Mei Dong, Junhao Zhu, Junwoo Jo

Estimating heterogeneous causal effect on networks via orthogonal learning

Yubai Yuan, The Pennsylvania State University

Abstract. Abstract: Many significant scientific research studies, such as those in epidemiology, and social sciences, involve estimating the effects of treatments, exposures, or interventions on populations where interference between units exists. The influence of one unit,Äôs treatment on other units, mediated through network interactions, is commonly referred to as spillover effects. Although the ubiquitous applications and theoretical significance, traditional causal inference methods cannot directly estimate spillover effects due to the interference over network. In addition, the spillover effect is typically heterogeneous in reality. Therefore, it is critical to appropriately specify the exposure mapping and its cross-unit variation. Inspired by the successful application of attention mechanisms in graph neural network, we propose a new method for causal inference on networks that enables researchers to estimate heterogeneous treatment and spillover effects. The proposed method provides interpretable structures for causal estimands and allows for the use of machine learning methods to estimate nuisance components in the models without relying on strong parametric assumptions. Compared to existing methods, the new framework employs a data-adaptive approach to estimate exposure mapping, rather than requiring the functional form to be known a priori. This is joint work Yuanchen Wu from Penn State.

Marginal Causal Effect Estimation with Continuous Instrumental Variables

Mei Dong, Division of Biostatistics, University of Toronto

Abstract. Instrumental variables (IVs) are often continuous, arising in diverse fields such as economics, epidemiology, and social sciences. Existing approaches for continuous IVs typically impose strong parametric models or assume homogeneous treatment effects, while fully nonparametric methods may perform poorly in moderate- to high-dimensional covariate settings. We propose a new framework for identifying the average treatment effect with continuous IVs via conditional weighted average derivative effects. Using a conditional Riesz representer, our framework unifies continuous and categorical IVs. In this framework, the average treatment effect is typically overidentified, leading to a semiparametric (observed-data) model$M_{sp}$with a nontrivial tangent space. Characterizing this tangent space involves a delicate construction of a second-order parametric submodel, which, to the best of our knowledge, has not been standard practice in the literature. For estimation and inference, building on an influence function that is not necessarily efficient underM_{sp}but is locally efficient within a submodel ofM_{sp}, we develop a locally efficient, triply robust, and easy-to-implement estimator. We apply our methods to an observational clinical study from the Princess Margaret Cancer Centre to examine the so-called obesity paradox in

oncology, assessing the causal effect of excess body weight on two-year mortality among patients with non-small cell lung cancer.

Modeling Cell Developmental Trajectory using Multinomial Unbalanced Optimal Transport

Junhao Zhu, Harvard University

Abstract. Inferring developmental trajectories from snapshot single-cell RNA sequencing (scRNA-seq) data remains a fundamental challenge due to the destructive nature of sequencing and the high level of noise in cellular measurements. While optimal transport (OT) methods applied at the single-cell level have gained increasing attention, they often suffer from high estimation variance and substantial computational cost. We propose a metacell based Multinomial Unbalanced OT framework to reconstruct cellular developmental dynamics. By aggregating similar cells into metacells prior to transport estimation, the proposed approach introduces structural constraints that improve both statistical stability and computational efficiency. Through analysis of a large-scale mouse developmental atlas data, we demonstrate that metacell-level OT provides more reliable estimates of transition probabilities and population growth rates than single-cell OT methods, which can be sensitive to technical variation and computational artifacts. Our results accurately recover cell-type transitions congruent with biological ground truth, suggesting that metacell aggregation is not merely a computational convenience but a statistical necessity for reliable OT trajectory inference. We further establish non-asymptotic and asymptotic convergence guarantees, as well as bootstrap consistency for valid confidence interval construction, under a finite-mixture model with data-driven transport costs.

Fully Bayesian synthetic control methods with sparse convex hull restriction and Gaussian process

Junwoo Jo, Kyungpook National University

Abstract. We propose a fully Bayesian synthetic control method that preserves the convex hull restriction of the conventional method while enabling coherent posterior inference. Our approach reparameterizes the weights by introducing latent positive variables and assigning them an exact spike-and-slab prior. This construction yields a Dirichlet prior on the normalized weights, automatically respects the convex hull restriction, and allows exact zeros, thereby inducing sparsity and enhancing interpretability. We develop a Gibbs sampling algorithm that jointly updates the inclusion indicators and latent weights, resolving the dimension-changing problem via a combination of Laplace approximation and population Monte Carlo algorithm. We model treatment effects with a Gaussian process prior to allow cross-time correlation. This prior flexibly captures dynamic effects and provides principled uncertainty quantification. In simulations based on a linear factor model with designed sparsity and outliers, the proposed method achieves the lowest mean squared error among the competing approaches. It also delivers higher interval coverage, particularly compared with Bayesian linear models using alternative priors. We further illustrate the practical advantages of our approach in an empirical application to California's tobacco control program.

Recent Advances in Structured High-Dimensional Learning (IS-75)

Time & Location: May 29, 08:45 AM - 10:25 AM | Room 108

Chair: Qishuo Yin

Proposer: Xingche Guo, University of Connecticut

Presenters: Qishuo Yin, Chandra Sekhar Dronavajjala, Reihaneh Malekian, Kevin Kapner

SMART Fine-tuning Factor Augmented Neural Lasso

Qishuo Yin, Princeton University

Abstract. Fine-tuning is a widely used strategy for adapting pre-trained models to new tasks, yet its methodology and theoretical properties in high-dimensional nonparametric settings with variable selection have not yet been developed. We propose a source-model-augmented residual tuning (SMART) framework, which incorporates the pre-trained source model as an augmented feature into the target learner and estimates only the residual target-specific component. The approach is widely applicable, from parametric and sparse models to neural networks and blackbox machine learning models. We focus on the development of fine-tuning factor-augmented neural Lasso, resulting in SMART-FAN-Lasso. This transfer-learning framework for high-dimensional nonparametric regression with variable selection simultaneously handles covariate and posterior shifts. We use a low-rank factor structure to manage high-dimensional dependent

covariates and a residual tuning decomposition in which the target function is expressed as a function of source model and other target-specific variables, thereby reducing the effective complexity of the target task. We derive minimax-optimal excess risk bounds, characterizing the precise conditions, in terms of relative sample sizes and function complexities, under which fine-tuning yields statistical acceleration over single-task learning. Extensive numerical experiments across diverse covariate- and posterior-shift scenarios demonstrate that SMART-FAN-Lasso consistently outperforms standard baselines and achieves near-oracle performance even under severe target sample size constraints, empirically validating the derived rates.

Does Your Prediction Set Actually Change the Decision? A Conformal Threshold for High-Dimensional Optimization

Chandra Sekhar Dronavajjala, University of Connecticut

Abstract. When an ML model's prediction feeds into an operational decision, such as dispatching energy generators or routing vehicles, the prediction is never perfect. Recent work in conformal prediction and robust optimization (Patel et al., AISTATS 2024; Chenreddy and Delage, UAI 2024) proposes wrapping these predictions in distribution-free uncertainty sets and letting the optimizer hedge against the uncertainty. The coverage guarantee ensures the true cost falls within the set at a target rate, say 90% of the time. This sounds like a safety net worth having. But before adopting it, a practitioner needs to know: does the uncertainty set actually change the decision, or does the optimizer pick the same action regardless? If the answer is "no change", the system gets a free coverage certificate. If "yes", the certificate comes at a cost, and the practitioner needs to budget for it. We show that the answer depends on the dimension of the problem and the combinatorial structure of the feasible set, not on the quality of the ML model. We identify a computable switching threshold that separates the two regimes: below the threshold, the decision is unchanged; above it, the optimizer switches to a more conservative alternative. On flow networks used in routing and dispatch, the threshold shrinks as the number of near-optimal alternative paths grows, which explains why low-dimensional problems like energy dispatch with 10 generators enjoy free coverage, while high-dimensional problems like taxi routing across 2,308 road segments see the decision change on nearly half the rounds. We validate this on four real-world benchmarks using public data from the California energy grid, NYC taxi records, and CDC flu surveillance. A predictor ablation confirms the key finding: improving the ML model's accuracy by 35% barely affects whether the decision changes, because it is the problem geometry, not the prediction error, that determines the threshold. We derive a single pre-deployment diagnostic from the stationary behavior of the online conformal calibration process that classifies any new problem instance into one of three regimes: free coverage, costly coverage, or a critical boundary between the two. generators or routing vehicles, the prediction is never perfect. Recent work in conformal prediction and robust optimization (Patel et al., AISTATS 2024; Chenreddy and Delage, UAI 2024) proposes wrapping these predictions in distribution-free uncertainty sets and letting the optimizer hedge against the uncertainty. The coverage guarantee ensures the true cost falls within the set at a target rate, say 90% of the time. This sounds like a safety net worth having. But before adopting it, a practitioner needs to know: does the uncertainty set actually change the decision, or does the optimizer pick the same action regardless? If the answer is ,Äúno change,,Äù the system gets a free coverage certificate. If ,Äúyes,,Äù the certificate comes at a cost, and the practitioner needs to budget for it. We show that the answer depends on the dimension of the problem and the combinatorial structure of the feasible set, not on the quality of the ML model. We identify a computable switching threshold that separates the two regimes: below the threshold, the decision is unchanged; above it, the optimizer switches to a more conservative alternative. On flow networks used in routing and dispatch, the threshold shrinks as the number of near-optimal alternative paths grows, which explains why low-dimensional problems like energy dispatch with 10 generators enjoy free coverage, while high-dimensional problems like taxi routing across 2,308 road segments see the decision change on nearly half the rounds. We validate this on four real-world benchmarks using public data from the California energy grid, NYC taxi records, and CDC flu surveillance. A predictor ablation confirms the key finding: improving the ML model,Äôs accuracy by 35% barely affects whether the decision changes, because it is the problem geometry, not the prediction error, that determines the threshold. We derive a single pre-deployment diagnostic from the stationary behavior of the online conformal calibration process that classifies any new problem instance into one of three regimes: free coverage,

costly coverage, or a critical boundary between the two.

LDP for Tensor Forms

Reihaneh Malekian, Columbia University

Abstract. In this paper, we study a large deviation principle (LDP) for a tensor-valued functional of i.i.d. random variables, when the sequence of tensors converges under a variant of the "bad" cut norm. By studying an LDP in this topology, we are able to improve conditions under which the LDP holds in applications of interest. As another application, we analyze a Gibbs measure with a tensor-valued Hamiltonian, and characterize the optimizers of the limiting variational problem in terms of a functional fixed point equation. We focus on several concrete examples, which include monochromatic subgraph counts in inhomogeneous random graphs, Erdős-R'enyi hypergraphs, and a generalized Potts model of order$v \geq 2$. Studying the optimization problem, we give sufficient conditions for uniqueness of the optimizer, as well as for existence of constant optimizers (replica symmetry). Our results demonstrate universal weak laws for a large class of tensor Gibbs models with approximately regular tensors.

Estimation of a rank-reduced sample covariate parameter matrix in generalized bilinear models

Kevin Kapner, Harvard University

Abstract. Generalized bilinear models (GBMs) provide a versatile framework for dimensionality reduction and effect estimation on high-dimensional non-Gaussian data, however these models are known to suffer from a decrease in statistical efficiency as the number of sample covariates increases. To address this fundamental issue, we developed a new model based on the traditional GBM framework, which we refer to as RR-GBM, which allows for the estimation of a reduced-rank version of the coefficient matrix for the sample covariates. This procedure improves estimation and uncertainty quantification when using a large number of sample covariates. When the true sample coefficient matrix is reduced rank or close to reduced rank, RR-GBM outperforms the standard GBM coefficient estimation procedure. Additionally, we provide a method for sample covariate selection, identification of the underlying rank of the true parameter matrix, and visualization of relationships among covariates and among features. We demonstrate the method in an application to scRNA-seq data.

Parallel Session 3 | 04:00 PM - 05:40 PM, May 28

Recent advancements of statistical methodologies in functional data analysis (IS-11)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 202

Chair: Hyemin Yeon

Proposer: Hyemin Yeon, Kent State University

Presenters: Sara Lopez-Pintado, Erjia Cui, Shanshan Wang, Hyemin Yeon

The Quantile Integrated Depth with Applications to Noisy Functional Data

Sara Lopez-Pintado, Northeastern University

Abstract. Functional data analysis involves data for which the basic unit of observation is a function or image. The development of robust exploratory tools and inferential methods is very much needed since few assumptions can be made about the generating process. Data depth, a well-known non-parametric tool for analyzing functional data, provides a rigorous method for ranking a sample of curves from the center outwards, allowing for robust inference and outlier detection. Several notions of depth for functional data have been introduced in the last few decades. Here we develop a new family of depths, termed quantile integrated depth (QID), that are based on integrating up to the$K$-th quantile of the univariate depths. We show that this new family of depths has desirable properties, including a type of invariance, maximality at the center, and monotonicity with respect to the deepest point. In addition, since functional data are commonly observed with noise, we explore the effect of noise on different notions of depth. Compared to alternatives, the proposed QID is shown to be robust and perform well on noisy functional data. We also illustrate the advantages of using$QID_K$to identify potential hard-to-detect shape outliers.

Sparse Longitudinal Functional Principal Components Analysis

Erjia Cui, University of Minnesota

Abstract. Accurately monitoring mental fatigue is critical for improving workplace safety and productivity. A recent study examined passively collected smartphone typing speed as an ambulatory metric of mental fatigue using data from the Intern Health Study (IHS). While average typing speed patterns were found to be consistent with validated measures of mental fatigue, how these trajectories vary across participants and across days within a participant remains an open question. Treating typing speed trajectories as sparsely observed functional data, we propose a novel sparse longitudinal functional principal components analysis (sparse LFPCA) method for decomposing variability and predicting individual curves. Specifically, sparse data are accommodated by casting covariance estimation as a structured penalized spline regression problem, enabling simultaneous estimation and smoothing of multiple covariance components while borrowing information across locations in the functional domain. Simulations show that sparse LFPCA (1) accurately estimates eigenfunctions and generates reasonable predictions for underlying curves, and (2) achieves similar or superior performance compared to existing methods for dense longitudinal and sparse multilevel functional data. By applying sparse LFPCA to the IHS data, we reveal new and interpretable participant- and day-level patterns not captured by previous analyses. The methods are accompanied by an R implementation, lfpca.sparse().

Nonlinear Sufficient Dimension Reduction for Conditional Quantiles in Scalar-on-Function Single-Index Models

Shanshan Wang, University of North Carolina at Charlotte

Abstract. Functional data analysis is crucial in many applications, yet its high-dimensional nature necessitates effective dimension reduction techniques. While existing approaches primarily focus on linear reductions, we introduce a nonlinear sufficient dimension reduction framework for conditional quantiles of single-index models when the predictors are random functions. Our approach constructs two nested functional spaces: a Hilbert space representing the functional data and a reproducing kernel Hilbert space that captures nonlinearity. The kernel in the latter is determined by the inner product of the former, leading to a natural hierarchical structure. We begin by characterizing dimension reduction at the general level ofσ-fields and proceed to that of classes of functions, leading to the notion of the central quantile class. We introduce our proposed estimator, called theτth functional generalized central quantile subspace (τ-fGCQS), and establish its convergence rate. Finally, we demonstrate the performance of our estimator through simulations and real-world applications to health studies, examining various health indicators, such as ADHD, Parkinson's disease, and BMI.

Gaussian and bootstrap approximations for functional principal component regression

Hyemin Yeon, Kent State University

Abstract. Asymptotic inference using functional principal component regression (FPCR) has long been considered difficult, largely because, upon any scalar scaling, the FPCR estimator fails to satisfy a central limit theorem, leading to the prevailing belief that it is unsuitable for direct statistical inference. In this paper, we upend this traditional viewpoint by establishing a new result: upon suitable operator scaling, valid Gaussian and bootstrap approximations hold for the FPCR estimator. We apply this surprising finding to hypothesis testing for the significance of the slope function in functional regression models and demonstrate the strong numerical performance of the resulting tests. While concise, our results yield powerful inferential tools for functional regression. We believe it paves the way for new lines of inferential methodology for more complex functional regression settings.

Career Development Panel in Statistics and Data Science (IS-15)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 101

Chair: Yang Liu

Proposer: Yang Liu, Upstart Network, Inc.

Panelists: Jieying Jiao, Nathan Lally, Xiaolin Chang, Erica McCullough, Yang Liu

New Advances in Time Series Analysis (IS-19)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 306

Chair: Yao Zheng

Proposer: Yao Zheng, University of Connecticut

Presenters: Kejin Wu, Zeda Li, Sumanta Basu, Han Xiao

Distributional Conformal Prediction for Markov Processes

Kejin Wu, Loyola University Chicago

Abstract. We extend the distributional conformal prediction method to a strictly stationary Markov process. Although there are conditional and marginal coverage guarantees on conformal prediction intervals with i.i.d data, the theoretical understanding of the coverage rate with Markov processes remains limited. In particular, our method explores the probability integral transform and the estimation inference of the corresponding CDF estimators with a Markov process. Instead of relying on a specific model structure to do predictions, the idea of the distributional conformal prediction interval aligns with the model-free prediction spirit. We build the non-asymptotic error bound of its unconditional coverage rate with beta-mixing condition. The asymptotic validity of the conditional prediction interval is also verified. In addition, we show that our conditional prediction interval is still asymptotically valid with Markov processes being L^p-m-approximable instead of satisfying the mixing property. The simulations and real data studies are deployed to compare our method with the Model-free Bootstrap method, which is an important model-free prediction alternative.

Mean Independent Component Analysis of Multivariate Time Series

Zeda Li, Baruch College, CUNY

Abstract. We introduce the mean independent component analysis for multivariate time series to reduce the parameter space. In particular, we seek for a contemporaneous linear transformation that detects univariate mean independent components so that each component can be modeled separately. The mean independent component analysis is flexible in the sense that no parametric model or distributional assumptions are made. We propose a unified framework to estimate the mean independent components from a data with a fixed dimension or a diverging dimension. We estimate the mean independent components by the martingale difference divergence so that the mean dependence across components and across time is minimized. The approach is extended to the group mean independent component analysis by imposing a group structure on the mean independent components. We further introduce a method to identify the group structure when it is unknown. The consistency of both proposed methods is established. Extensive simulations and a real data illustration for community mobility is provided to demonstrate the efficacy of our method.

Autotune: fast, accurate, and automatic tuning parameter selection for Lasso

Sumanta Basu, Cornell Uniersity

Abstract. Least absolute shrinkage and selection operator (Lasso), a popular method for high-dimensional regression, is now used widely for estimating high-dimensional time series models such as the vector autoregression (VAR). Selecting its tuning parameter efficiently and accurately remains a challenge, despite the abundance of available methods for doing so. We propose autotune, a strategy for Lasso to automatically tune itself by alternately estimating regression coefficients and noise standard deviation. Simulation experiments on regression and VAR models show that autotune is faster than alternative methods, and more accurate when signal-to-noise ratio is low. It also offers a new estimator of noise scale and new diagnostic plots to check model sparsity. Finally, we demonstrate the utility of autotune on a real-world financial data set and provide an R package based on C++ on GitHub.

Dynamic matrix factor model for counts data

Han Xiao, Rutgers University

Abstract. A dynamic factor model is proposed for matrix time series, where the observations are counts. The model is formulated as Poisson observations conditional on the rate matrices, which have log-normal distributions. The logarithm of the rate matrices has the form of a dynamic Gaussian factor model of matrix time series, where the dynamics are captured by the factor process. We use a log moment method to estimate the loading matrices, and the variational inference to estimate the factor process. An autoregressive model is imposed on the factor process to enable predictions. It is also considered to include a trend component in the log rate matrices to capture the possible nonstationarity. Theoretical and numerical analyses are conducted for the proposed model.

Advancement in transfer learning and high dimensional inference (IS-20)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 305

Chair: Qingkai Dong

Proposer: Wen Zhou, New York University

Presenters: Qian Tang, Zifeng Zhang, Yu Gui, Jinhang Chai

A Robust Transfer Learning Framework for High-Dimensional Multiclass Distance Weighted Discrimination

Qian Tang, University of Minnesota

Abstract. Multiclass classification remains a challenging problem, particularly in high-dimensional settings where the number of features can greatly exceed the sample size. To address this challenge, we develop a transfer learning framework for multiclass distance weighted discrimination (DWD) that leverages auxiliary information from related source domains to improve target-domain classification. To mitigate negative transfer, we further propose an adaptive strategy that assesses source-target similarity and regulates the contribution of auxiliary data, thereby retaining beneficial source information while reducing the impact of irrelevant or harmful sources. Extensive simulation studies and real data applications demonstrate the effectiveness and robustness of the proposed method.

Dual Validity of Robust Standard Errors under Complex Dependence

Zifeng Zhang, Yale School of Public Health

Abstract. Regression is a workhorse for analyzing data from randomized experiments, adaptive trials, and related designs, where valid inference typically relies on robust standard errors. These adjustments are usually justified as correcting dependence in unobserved errors, such as dependence induced by shocks, interference, or hidden common causes, even though such dependence is rarely observable or reliably modeled in practice. By contrast, the dependence structure of the regressor of interest, most notably treatment assignment, is often known, testable, and engineered by design. Motivated by this asymmetry, we develop a general asymptotic framework for regression inference under unknown and potentially unstructured dependence and establish a dual validity principle: various robust t-statistics remain asymptotically valid if the correlation structure encoded by the chosen standard-error adjustment is correct for either the regressor of interest or the unobserved errors, even when the other component is misspecified. Establishing this predictor-error symmetry is mathematically nontrivial and, to our knowledge, has not been formalized in the classical robust-inference literature despite its centrality to empirical practice. For example, cluster-robust inference remains valid when either errors are clustered or assignment is clustered. The result reframes robust standard errors as design-aligned devices: when error dependence is uncertain, robust adjustments should be chosen to match the known assignment correlation, rather than used as a catch-all correction for unspecified error structure.

Distributionally robust risk evaluation with an isotonic constraint

Yu Gui, University of Pennsylvania

Abstract. Statistical learning under distribution shift is challenging when neither prior knowledge nor data from the target distribution is available. Distributionally robust learning (DRL) aims to control the worst-case statistical performance within a set of candidate distributions, but how to properly specify the set remains challenging. To enable distributional robustness without being overly conservative, in this paper we propose a shape-constrained approach to DRL, which incorporates prior information about the way in which the unknown target distribution differs from its estimate—specifically, we assume the unknown density ratio between the target distribution and its estimate is isotonic with respect to some partial order. At the population level, we provide a solution to the shape-constrained optimization problem that can be solved without the challenge of an explicit isotonic constraint. At the sample level, we provide consistency results for an empirical estimator of the target in a range of different settings. Empirical studies on both synthetic and real data demonstrate the improved efficiency of the proposed shape-constrained approach.

Deep Transfer Q-Learning for Offline Non-Stationary Reinforcement Learning

Jinhang Chai, Princeton University

Abstract. In dynamic decision-making scenarios across business and healthcare, leveraging sample trajectories from diverse populations can significantly enhance reinforcement learning (RL) performance for specific target populations, especially when sample sizes are limited. While existing transfer learning methods primarily focus on linear regression settings, they lack direct applicability to reinforcement learning algorithms. This paper pioneers the study of transfer learning for dynamic decision scenarios modeled by non-stationary

finite-horizon Markov decision processes, utilizing neural networks as powerful function approximators and backward inductive learning. We demonstrate that naive sample pooling strategies, effective in regression settings, fail in Markov decision processes. To address this challenge, we introduce a novel "re-weighted targeting procedure" to construct "transferable RL samples" and propose "transfer deep Q-learning", enabling neural network approximation with theoretical guarantees. We assume that the reward functions are transferable and deal with both situations in which the transition densities are transferable or non-transferable. Our analytical techniques for transfer learning in neural network approximation and transition density transfers have broader implications, extending to supervised transfer learning with neural networks and domain shift scenarios. Empirical experiments on both synthetic and real datasets corroborate the advantages of our method, showcasing its potential for improving decision-making through strategically constructing transferable RL samples in non-stationary reinforcement learning contexts.

Statistical Research at Bentley University: Density Power Divergence, Model-Assisted Estimation, Casual Effects of Mandatory Testing, and Interest-Rate Modeling (IS-32)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 206

Chair: Jackson Lautier

Proposer: Jackson Lautier, Bentley University

Presenters: Tony Ng, Reagan Mozer, Edward J Kim, Reuben Brefo Marfo

Discrimination Between Generalized Exponential and Weibull Distributions Using a Density Power Divergence Measure

Hon Keung Tony Ng, Department of Mathematical Sciences, Bentley University

Abstract. Discriminating between two similar candidate statistical models for a given data set based on the conventional ratio of maximized likelihoods has been explored in great detail in the literature. The problem of model discrimination becomes more complicated when the candidate models are non-nested, possess the same number of parameters, and resemble each other closely for a certain range in the parametric space, with only a handful of different characteristics that are difficult to extract or identify from a given data set. The conventional method may fail to provide conclusive discriminatory evidence toward either model for such cases. A discrimination criterion based on the density power divergence will be introduced for model discrimination between the generalized exponential distribution and the Weibull distribution, along with a brief review of the existing method. Some theoretical properties of the proposed method are discussed. A Monte Carlo simulation study is used to evaluate the performance of the proposed method and compare it to the existing method under different scenarios. The utility and applicability of the proposed method are illustrated through a real data set.

Stratified Sampling for Model-Assisted Estimation with Surrogate Outcomes

Reagan Mozer, Bentley University

Abstract. In many randomized trials, outcomes such as essays or open-ended responses must be manually scored before impact analysis, a process that is costly and limiting. Model-assisted estimation combines surrogate outcomes from machine learning or large language models with a human-coded subset to obtain unbiased estimates, but existing approaches rely on simple random sampling and ignore systematic structure in prediction errors. We extend this framework by incorporating stratified sampling to more efficiently allocate human coding effort. We derive the exact variance of the stratified estimator, characterize conditions under which stratification improves precision, and identify a Neyman-type optimal allocation rule that oversamples strata with larger residual variance. Comprehensive simulation studies confirm that stratification consistently improves efficiency when surrogate prediction errors exhibit structured bias or heteroskedasticity. We present two empirical applications, including an education RCT and a large observational corpus, to illustrate practical implementation using ChatGPT-generated surrogate outcomes. Overall, this framework provides a practical design-based approach for leveraging surrogate outcomes and strategically allocating human coding effort to obtain unbiased estimates with greater efficiency. While motivated by text-as-data applications, the methodology applies broadly to any setting where outcome measurement is costly, including both group comparisons and single-group mean estimation.

The Causal Effect of Mandatory Testing Policies on Private Tutoring Markets

Edward J Kim, Bentley University

Abstract. Multiple studies suggest that policies mandating college entrance exams improve college outcomes, especially for students who would otherwise not sit for the exam. Less understood is how this increased competition for college admissions affects the marketplace for other college preparatory resources. Our study estimates that such statewide mandatory testing policies increase the prevalence of private tutoring centers by 16.1% compared to the counterfactual, with more pronounced effects in high income, highly educated, and high proportion Asian areas. The results were robust to model specification choices, subsample analyses, and placebo tests using music instruction and martial arts industries could not replicate such patterns. Our findings suggest that interventions to further educational equality may have second order effects that undermine their impact as the private sector responds to allow resourced families to stay competitive.

Forecasting Interest Rates Through Data Partitioning: Evidence from Developed and Emerging Markets

Reuben Brefo Marfo, Bentley University

Abstract. This study examines the forecasting performance of short-rate models for yield-to-maturity curves across a diverse range of economies. Building on the approach developed by Orlando, Mininni, and Bufalo (2019) [1], we apply the Cox–Ingersoll–Ross (CIR) [2] and Vasicek [3] models to weekly yield-to-maturity data at multiple maturities (3M, 6M, 1Y, 5Y, and 10Y). Our dataset encompasses developed countries, including Canada, France, Germany, Italy, Japan, Portugal, Singapore, Switzerland, the United Kingdom, and the USA, as well as emerging economies such as Bangladesh, Brazil, Egypt, India, Indonesia, Côte d'Ivoire, Mexico, the Philippines, Sri Lanka, and Uganda. Following the framework outlined in [1], each yield series is segmented into statistically homogeneous subsamples by testing for both normal and noncentral chi-square distributions. Within each segment, the models are calibrated using closed-form estimators, and one-step-ahead forecasts are generated through the corresponding conditional expectations. This methodology enables the detection of structural breaks and regime shifts, allowing the models to capture dynamic changes in volatility and mean reversion that differ markedly across economic environments. We demonstrate that emerging economies typically exhibit higher long-term mean yields, slower speeds of mean reversion, and greater volatility compared to developed markets. Moreover, the segmentation approach—particularly when using noncentral chi-square partitioning—improves forecasting accuracy relative to traditional methods, underscoring the benefit of employing high-frequency data to capture rapid market shifts. In addition, this study extends prior work by incorporating EWMA-based benchmarks and conducting statistical comparisons between models, including paired t-tests, regression analysis on performance differentials across country groups, and repeated-measures (linear mixed) models to evaluate differences in forecasting performance between developed and emerging economies. By extending Orlando et al.'s approach to a diverse global dataset and multiple maturities, our work contributes a unified framework for yield curve forecasting in both emerging and developed economies. These findings have significant implications for policymakers and fixed-income investors in managing risk and making informed investment decisions.

Modern Statistical Inference for Complex Data and Adaptive Experiments (IS-41)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 301

Chair: Buddika Peiris

Proposer: Buddika Peiris, Worcester Polytechnic Institute

Presenters: Adam Sales, Hammed Olayinka, Saad Mouti, Buddika Peiris

Exact Fisherian P-Values for Multi-Armed Bandits

Adam Sales, Worcester Polytechnic Institute

Abstract. Randomized multi-armed bandit experimental designs in which successive subjects are adaptively randomized between conditions: randomization probabilities are based on previous users,Äô outcomes. One heuristic, Thompson sampling, essentially randomizes users to conditions according to the posterior probabilities that each condition is optimal, conditional on previous users,Äô responses. However, from a scientific perspective, RMABs

present a challenge: the adaptivity of RMABs induces a complex dependence structure between the observations, which invalidates usual approaches to statistical inference from randomized experiments, which assume some degree of independence between observations. This paper illustrates a simple, but overlooked, solution: simulation-based exact p-values for Fisher,Äôs strict null hypothesis of no effect. This approach requires a complete record of randomization probabilities, treatment assignments, and outcomes,Äîand sufficient computational power,Äîbut little else. We illustrate this approach using a new dataset of over 200 RMABs conducted on an online homework platform, where the outcome of interest was students,Äô correctness on the next problem, and compare randomization-based p-values using a variety of test statistics to na√Øve chi-squared tests.

Bayesian Predictive Inference for Multiple Series with Correlated Spatial Priors on Autoregressive Parameters

Hammed Abiola Olayinka, Worcester Polytechnic Institute

Abstract. We present a Bayesian multiple time-series model for panel data that improves forecasting by borrowing information across neighboring areas. Each unit is modeled with a simple univariate autoregressive likelihood, while spatial dependence is introduced entirely through a conditional autoregressive prior placed on unit-level intercept and autoregressive coefficient deviations. Inference is conducted using a collapsed Gibbs sampler, which reduces posterior dependences, improves mixing and effective sample size, and preserves interpretable unit-level dynamics. This design avoids the overparameterization of multivariate formulations while enabling spatially targeted pooling that respects geographic adjacency. We derive stable posterior updates under standard positive-definiteness conditions for the conditional autoregressive precision. Empirically, we study an annual panel of average hourly earnings for production employees across selected California metropolitan statistical areas and compare the proposed model to independent single-series autoregressive models and a pooled multiple time-series benchmark without spatial structure. The proposed approach delivers consistent gains in point accuracy and interval calibration, providing a flexible and interpretable framework for forecasting and small-area estimation in regional economic analysis and related domains.

Inference on Cross-Sectional Fit in Linear Factor Models

Saad Mouti, University of New Haven

Abstract. This talk develops inference for a cross-sectional R^2 constructed from estimated intercepts in multivariate time-series regressions and for differences in this measure across competing linear factor models. In traded-factor asset pricing, the statistic measures the fraction of cross-sectional dispersion in mean returns explained by the model. We show that its asymptotic behavior is regime dependent: under interior configurations it is T^0.5-Gaussian with feasible HAC variance estimation, whereas at perfect fit and related boundary cases it is nonregular and converges at rate T to quadratic-form limits. We develop feasible inference for each regime and comparison tests for nested and non-nested models. Simulations and empirical applications show that apparent differences in model fit often reflect substantial sampling uncertainty.

Synthesizing ROC Curves

Buddika Peiris, Worcester Polytechnic Institute

Abstract. In this work we propose an improved meta-analysis method to synthesize the ROC curves from multiple individual studies. When synthesizing, ROC curves are transformed to linear models then combined and re-transformed to a ROC curve. The existing method proposed in Kester and Buntinx (2000) has a serious issue since information about the covariance between the coefficients of the same model and covariance among coefficients across studies are not utilized. Here we introduce how to take the correlations between the coefficients and the correlations between studies into account to improve the existing meta-analysis. We provide numerical example to compare the proposed methods with the existing method by using mean square prediction errors with applications to forecasting problem in Environmental study. Keywords: ROC Curve, meta-analysis, generalized least square, Der Simonian and Laird, synthesis of slopes.

Modern Statistical Learning for Complex Data: Methods and Applications (IS-44)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 201

Chair: Yeongjin Gwon

Proposer: Yeongjin Gwon, University of Nebraska Medical Center

Presenters: Vishal Midya, Qi Zhang, Fan Dai, Inkoo Lee

Risk stratification model for autism using machine-learning analysis of molecular temporal dynamics in hair - a multicenter study

Vishal Midya, Icahn School of Medicine at Mount Sinai

Abstract. Absence of autism risk-stratification tools under 18 months hampers early intervention. In a multi-national sample of 1697 (from California (n= 1112), New York City (n= 123), Sweden (n= 306), Japan (n= 110), and Mexico City (n= 46), with 97% below 21 years-of-age), autism was assessed using DSM-5 criteria for autism spectrum disorder or gold standard diagnostic instruments (ADOS-2 and/or ADI-R). A single hair strand from children aged 1 month or older was analyzed using laser ablation inductively coupled plasma mass spectrometry to sample along the shaft, generating time-series data at ~800 time points for 12 elemental intensities. We developed a first-stage model to stratify individuals as a low autism probability group and extended this approach to a sequential triage framework by training and applying a second-stage model to participants not classified as low probability, thereby further stratifying them into intermediate and high autism probability groups. Models were trained, ensembled, and tuned on participants from California and Sweden, then tested on 580 participants (within-population replication and external population testing in New York, Mexico, and Japan). Likelihood ratios (95%CI) for autism in low-, intermediate-, and high-probability groups were 0.18(0.15-0.23), 1.09(0.99-1.20), and 2.62(1.55-4.00), respectively. Low-probability classification (first-stage) had sensitivity of 96%(0.91-0.98), and high-probability classification (second-stage) had specificity of 90%(0.86-0.92). Assignment to progressively higher ASD risk categories was associated with increased odds of autism diagnosis (OR[95%CI]: 2.41[1.56,3.72]). We show that analyzing elemental biodynamics using single hair strands may provide an objective approach to stratifying autism likelihood in early childhood.

High dimensional mediation analysis with non-gaussian outcomes

Qi Zhang, University of New Hampshire

Abstract. High-dimensional mediation analysis has attracted increasing attention in recent years, motivated by applications in genomics and biomedical imaging where both exposures and mediators may be high-dimensional. Existing methods largely focus on mediator selection or rely on linear mediation models with continuous outcomes, which limits applicability in high-dimensional multi-omics studies and settings with non-Gaussian responses. Recent work by Zhang et al. (2024) introduced Mediation via Difference-in-Coefficients (MedDiC) framework for estimating causal effects in high-dimensional linear mediation models using debiased LASSO techniques. In this project, we extends MedDiC to generalized linear models.

Latent Factor Modeling for High-dimensional Data on the Unit Sphere

Fan Dai, Department of Statistics, North Dakota State University

Abstract. High-dimensional data on the unit sphere arise frequently across disciplines, either naturally or as a result of preprocessing such as normalization, and often display complex dependence structures. We develop an exploratory factor analysis framework for the projected normal distribution to explain such variability through a small number of interpretable latent factors. The proposed methodology is estimated by maximum likelihood via a fast, novel alternating expectation profile conditional maximization algorithm. Simulation studies show uniformly strong performance over a wide range of settings. We further demonstrate the practical value of the method through applications to tweets containing the #MeToo hashtag from early December 2018, time-course functional magnetic resonance images of the average pre-teen brain at rest, handwritten digits, and gene expression data from cancer cells in The Cancer Genome Atlas.

A Bayesian joint model with multivariate skew-t distribution for informative cluster size

Inkoo Lee, University of Georgia

Abstract. Pocket depth (PD) and clinical attachment loss (CAL) are important clinical measurements to assess the severity of periodontal disease. Previous studies have used the multivariate skewed-t distribution to model PD and CAL together, assuming data are missing at random. However, the assumption of data missing at random may be violated because PD and CAL are related to the number of teeth

within a person, resulting in the issue of informative cluster size. As the measurements are clustered within teeth, we extended the multivariate skewed-t model to a joint model using a Bayesian estimation approach. The multivariate skewed-t model was reconstructed in a hierarchical structure, and the cluster size model assumed a binomial distribution with a logit link. We propose a joint shrinkage prior for regression coefficients that induces row-wise sparsity across sites, while incorporating site-specific local shrinkage to allow column-wise sparsity within covariates. The proposed model was evaluated in simulations and compared to the multivariate skewed-t model without the cluster size. We then applied the proposed joint model to dental data in the San Juan Overweight Adults Longitudinal Study.

Quantifying Risk: Recent Developments in Probability of Technical Success Assessment for Pharmaceutical R&D Pipeline (IS-47)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 205

Chair: Steven A Gilbert

Proposer: Satrajit Roychoudhury, Pfizer Inc

Presenters: Adam Brown, Angela Zhu, Anindita Banerjee, Wei Wei

A Multimodal Causal AI Framework for Clinical Development PRTS Estimation

Adam Brown, PhaseV

Abstract. Predicting clinical trial outcomes remains a primary challenge in drug development, as traditional methods often rely on coarse historical baselines or non-causal associations. Single-model approaches frequently fail to capture the complex, multi-dimensional risks inherent in asset-specific programs. To address these limitations, PhaseV has developed a sophisticated Causal AI Platform designed to provide high-resolution decision support. The platform synthesizes heterogeneous evidence and channels causal insights into a unified estimate of Probability of Technical Success (PoTS) and risk-adjusted Net Present Value (aNPV). The PhaseV architecture utilizes a three-layer Bayesian synthesis framework to integrate diverse sources of causally-informative information: Layer 1 (Historical Precedent): Establishes indication- and modality-specific priors across all previously run trials. Layer 2 (Predictive Machine Learning): Captures empirical patterns of ~80 trial features in a deeply curated database consisting of thousands of studies. Layer 3 (Mechanistic/Statistical Modeling): Calculates PRTS from first principles, leveraging best-in-class ML models where needed to refine parameter estimates like effect size, operational characteristics, and power The framework specifically addresses model redundancy through a learned covariance structure, preventing "double-counting" of correlated signals. We add to this quantitative estimate of PRTS a qualitative layer, Layer 4 (Structured Failure-Mode Analysis), which decomposes risk into interpretable pathways (e.g., population variability, placebo inflation, or operational execution) and highlights potential risks for a given program, design, or indication. Lastly, we leverage the estimated PRTS and potential failure modes to calculate a risk-adjusted Net Present Value (aNPV), which is key for understanding potential return on investment before embarking on clinical studies. Taken together, our approach goes beyond traditional, single-model approaches to effectively account for multiple types of causal and correlative signals while maintaining explainability for a clinical and portfolio strategy audience.

From endpoints to outcomes: Integrating endpoint-level evidence into PTRS

Angela Zhu, Boehringer Ingelheim

Abstract. Interest in quantitative frameworks to estimate the probability of success in drug development has grown due to such factors as rising costs and the need for more informed portfolio decision-making. This presentation will provide an overview of the probability of technical and regulatory success (PTRS) framework based on observing a positive proof-of-concept outcome, highlighting the data inputs and integration of evidence across various stages of development. We then depict a practical implementation in systemic lupus erythematosus (SLE) to support clinical development planning, focusing on deriving the technical success component from key clinical endpoints. In this case study, we demonstrate how incorporation of historical benchmarking and understanding from earlier phases inform the construction of prior distributions for parameters of interest for a binary outcome. We conclude with a discussion of broader considerations and ongoing challenges in PTRS estimation, including evolving regulatory expectations, dose finding, and endpoint definitions. We highlight the importance of addressing sources of uncertainty moving from proof-of-concept to confirmatory phases and provide

a practical perspective on improving PTRS assessment to support decision-making.

Repro Samples and Sampling-based Inference (IS-56)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 111

Chair: Jiaqi Liu

Proposer: HaiYing Wang, University of Connecticut

Presenters: Minge Xie, Junyi Li, Dan Kluger, HaiYing Wang

Model-Free Inference for High-Dimensional Binary Classification Using Repro Samples

Minge Xie, Rutgers University

Abstract. In high-dimensional settings, when the underlying data-generating regression structure is completely unknown and unspecified, conducting statistical inference is a particularly challenging task. In this talk, we develop a model-free inference approach for high-dimensional binary classification, leveraging a class of (likely misspecified) sparsity generalized models. Our method builds on the so-called repro samples framework, which utilizes artificial samples that are generated by mimicking the true data or noises. The proposed approach facilitates inference by targeting both the model support and the regression coefficients of the oracle model, defined as the working model closest to the unknown true model. The method offers three key advantages: (1) it is fully model-free, requiring neither correct model specification nor sparsity of the true model; (2) it constructs a candidate set of influential covariates with guaranteed coverage under weak signal conditions; and (3) it provides confidence sets for any linear transformation of the oracle coefficients. When the oracle sparse model coincides with the true underlying model, the inference results directly apply to the true model, allowing simultaneous quantification of uncertainty in (discrete) model selection and (continuous) parameter estimation. Simulation studies illustrate the effective performance of the proposed method. An application to single-cell RNA-seq immune response data demonstrates that it identifies key genes, offering new insights into cellular immune response mechanisms.

Change Point Detection Inference with Unknown Number of Change Points Using Repro-Samples

Junyi Li, Rutgers, The State University of New Jersey

Abstract. In this talk, we develop a novel inference approach for change-point detection and uncertainty quantification when the number of change points is unknown for uni- and multivariate data. Our approach is developed based on the repro-samples framework, which constructs inference procedures through artificially generated samples or noise that mimic the observed data. When the data-generating distribution belongs to an exponential family, the proposed method achieves exact finite-sample coverage by conditioning on sufficient statistics. When the form of the distribution is unknown, the same approach remains applicable and yields asymptotically valid inference. To address the uncertainty in the unknown number of change points and to reduce the search space, we propose a candidate set construction approach using Fisher's inversion technique. We show that even under weak signal strength condition, the candidate set constructed can contain the true number of change points with high probability. Overall, the proposed approach provides a unified and flexible framework for valid inference in change-point detection while accounting for uncertainty in the number of change points. Simulation studies and real data analyses on both univariate and multivariate datasets are used to demonstrate the effectiveness of the proposed method, including applications to human activity data and sequential image data such as MNIST and CIFAR, where the goal is to detect changes in the underlying state or class over time.

M-estimation under Two-Phase Multiwave Sampling with Applications to Prediction-Powered Inference

Dan Kluger, Massachusetts Institute of Technology

Abstract. In two-phase multiwave sampling, inexpensive measurements are first collected on a large sample, and subsequently, expensive measurements are adaptively obtained on subsets of units across multiple data collection waves. Adaptively collecting the expensive measurements can increase efficiency but complicates statistical inference. We give valid estimators and confidence intervals for M-estimation under adaptive two-phase multiwave sampling. We focus on the case where proxies for the expensive variables—such as predictions from pretrained ma-

chine learning models—are available for all units and propose a Multiwave Predict-Then-Debias estimator that combines proxy information with the expensive, higher-quality measurements to improve efficiency while removing bias. We establish asymptotic linearity and normality and propose asymptotically valid confidence intervals. We also develop an approximately greedy sampling strategy that improves efficiency relative to uniform sampling. Data-based simulation studies support the theoretical results and demonstrate efficiency gains.

Two Types of Parameters in Subsampling for Massive Data

HaiYing Wang, UConn

Abstract. As massive datasets become the norm, subsampling has emerged as a crucial technique to make statistical computation feasible. This talk introduces the data-dependent subsampling approach, focusing on the critical distinction between two types of targets in subsampling: approximating the full-data estimator versus estimating the true population parameter. When the goal is to approximate a computationally intractable full-data estimator, conditional distributions are sufficient, and Inverse Probability Weighting (IPW) combined with optimal design (e.g., OSMAC) provides a robust solution, even under model misspecification. Conversely, when the goal is to infer the true population parameter, the unconditional distribution of the data becomes relevant, and unweighted approaches,ÄîsuchÄîas likelihood-based methods,ÄîoftenÄîyield significantly higher statistical efficiency, provided the model is correctly specified. We will discuss the theoretical properties of both approaches and explore their connections and differences. We will also point out some interesting facts regarding subsampling.

Recent Advances in Spatial Statistics and Complex Distributional Modeling (IS-78)

Time & Location: May 28, 04:00 PM - 05:40 PM | Room 108

Chair: Yu Wang

Proposer: Xingche Guo, University of Connecticut

Presenters: Yu Wang, Haewon Hwang, Manushi Siriwardana, Souhardya Sengupta

Robust Joint Modeling for Data with Continuous and Binary Responses

Yu Wang, University of Massachusetts Amherst

Abstract. In many supervised learning applications, the response consists of both continuous and binary outcomes. Studies have shown that jointly modeling such mixed-type responses can substantially improve predictive performance compared to separate analyses. However, outliers pose a new challenge to the existing likelihood-based modeling approaches. In this paper, we propose a new robust joint modeling framework for data with both continuous and binary responses. It is based on the density power divergence (DPD) loss function with the L1 regularization. The proposed framework leads to a sparse estimator that simultaneously predicts continuous and binary responses in high-dimensional input settings while down-weighting influential outliers and mislabeled samples. We also develop an efficient proximal gradient algorithm with Barzilai-Borwein spectral step size and a robust information criterion (RIC) for data-driven selection of the penalty parameters. Extensive simulation studies under a variety of contamination schemes demonstrate that the proposed method achieves lower prediction error and more accurate parameter estimation than several competing approaches. A real case study on wafer lapping in semiconductor manufacturing further illustrates the practical gains in predictive accuracy, robustness, and interpretability of the proposed framework.

On Kibble-Downton-Type Bivariate Distributions

Haewon Hwang, University of New Hampshire

Abstract. In this paper, we present a unified Kibble,ÄìDownton,Äìtype construction for bivariate lifetime models arising from a successive-shock (successive-damage) mechanism. Building on the classical Moran-Downton bivariate exponential and Kibble bivariate gamma distributions, we outline three general directions for extension and develop two representative inverse Gaussian-based models: the Moran-Downton bivariate inverse Gaussian (MDBIG) and the Kibble bivariate inverse Gaussian (KBIG) distributions. For each model, we study basic distributional properties, provide straightforward algorithms for generating random variates, and develop maximum-likelihood estimation procedures. Finite-sample performance is assessed via Monte Carlo simulation, and the methodology is illustrated using the electronic treeing data.

Mixture-based Nonparametric Estimation of Spatial Covariance Functions with Applications to HIV Key Population Size Estimation across Sub-Saharan Africa

Manushi Siriwardana, Pennsylvania State University

Abstract. Consistent data on the sizes of key populations, such as female sex workers (FSWs), are often scarce, particularly at the sub-national level. Accurate size estimates are critical to effectively allocate resources and achieve HIV targets. Since FSW population sizes may be spatially correlated across areas, models that account for spatial dependence can improve estimation. An important component of such models is the covariance function, which characterizes the spatial dependence structure of the underlying process. In this work, we study spatial covariance functions to estimate FSW population sizes in Sub-Saharan Africa (SSA). Many spatial models rely on parametric covariance functions. However, parametric estimation can suffer from model mis-specification, potentially leading to inefficient or biased predictions. We therefore develop a robust non-parametric approach for estimating the covariance function of a stationary isotropic process in$\mathbb{R}^d$. We focus on a class of covariance functions that are valid in all dimensions, which includes popular kernels such as the exponential and Mat'{e}rn kernels. Leveraging the fact that such covariance functions can be represented as infinite mixtures of scaled Gaussian kernels, we propose two estimation methods: weighted least squares and nonparametric maximum likelihood estimation to estimate the mixing measure of scaled Gaussian kernels. We also develop computationally efficient methods to solve these optimization problems using non-negative least squares and second-order descent updates. We evaluate the proposed methods through simulations and apply them to estimate the FSW population sizes at the sub-national level in SSA.

Theℓ-test: leveraging sparsity in the Gaussian linear model for improved inference

Souhardya Sengupta, Harvard University

Abstract. We develop novel LASSO-based methods for coefficient testing and confidence interval construction in the Gaussian linear model with$n \geq d$. Our methods' finite-sample validity is identical to that of their ubiquitous ordinary-least-squares-t-test-based analogues, yet have substantially higher power when the true coefficient vector is sparse. In particular, under sparsity our coefficient test, which we call theℓ-test, performs like the *one-sided*t-test (despite not being given any information about the sign), andℓ-test-based confidence intervals are correspondingly shorter than the standard$t$-test-based intervals. The nature of theℓ-test directly provides a novel exact adjustment conditional on LASSO selection for post-selection inference, allowing for the construction of post-selection$p$-values and confidence intervals. None of our methods require resampling or Monte Carlo estimation. We perform a variety of simulations and a real data analysis on an HIV drug resistance data set to demonstrate the benefits of theℓ-test. We additionally show that theℓ-test can be applied to a large class of asymptotically Gaussian estimators, dramatically expanding its applicability beyond linear models. This is a joint work with Lucas Janson.

Parallel Session 5 | 01:30 PM - 03:10 PM, May 29

Leveraging Statistical Learning for Sustainable Business and Industrial Solutions (IS-12)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 201

Chair: Kelum Gajamannage

Proposer: Kelum Gajamannage, Department of Mathematics and Applied Mathematical Sciences, University of Rhode Island

Presenters: Randy Paffenroth, Andres Torres, Nhu Nguyen, Kelum Gajamannage

Machine Learning and Industry 4.0

Randy Paffenroth, Worcester Polytechnic Institute

Abstract. Machine learning has had an impact on many aspects of modern life, including self-driving cars, modern cell phones, and web marketing, to name but a few. However, domains in which Machine Learning, and related techniques, hold the promise to transform large swaths of society and the economy are in their application to engineering, manufacturing, and the physical sciences. Such methods vary in complexity from the simple (e.g., linear regression) to the complicated (e.g., Neural Networks), and while quite powerful, they can also be difficult to use effectively. It is interesting to note that there are surprisingly many open questions on how such techniques work, especially considering

how much we depend on them every day, and in this talk we will discuss the relationship between modern Machine Learning ideas and Industry 4.0.

Some Applications of New Results in Stochastic Approximation with Discontinuous Drifts

Nhu N. Nguyen, University of Rhode Island

Abstract. The paper is concerned with stochastic approximation algorithms. Our main effort is focused on recently developed set-valued stochastic approximation methods. We begin with a brief introduction on stochastic approximation. Next, recent results are reviewed. Then the rest of the paper concentrates on applications of stochastic applications of set-valued problems.

Real-time forecasting of time series in financial markets using sequentially trained dual-LSTMs

Kelum Gajamannage, University of Rhode Island

Abstract. Financial markets are highly complex and volatile; thus, accurate forecasting of such markets is vital to make early alerts about crashes and subsequent recoveries. People have been using learning tools from diverse fields such as financial mathematics and machine learning to make trustworthy forecasts on such markets. However, the accuracy of such techniques had not been adequate until artificial neural network frameworks such as long short-term memory (LSTM) were utilized. Moreover, making accurate real-time forecasting, also known as nowcasting, of financial time series is highly sensitive to the LSTM's architecture in use and the procedure of training it. Herein, we forecast financial markets in real-time by training a dual version of LSTM, which forecasts only one time step at each iteration so that the forecast for this iteration will be in the input for the next iteration. Semi-convergence is a prominent issue in a recurrent LSTM setup as the error could propagate through iterations; however, the duality of this LSTM aids in dwindling this issue. Especially, we employ one LSTM to find the best number of epochs associated with the least loss and train the second LSTM only through that many epochs to make forecasting. We treat the current forecast as a part of the training set for the next forecast and train the same LSTM. While classic ways of training cause more error when the forecast is made further away through the test period, our approach offers superior accuracy as the training increases when it proceeds through the testing period. The forecasting accuracy of our approach is validated using three time series from each of the three diverse financial markets: stock, cryptocurrency, and commodity. The results are compared with those of a single LSTM, an extended Kalman filter, and an autoregressive integrated moving average model.

Recent advances in survey sampling research (IS-17)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 302

Chair: Jing Wang

Proposer: Jae-kwang Kim, Iowa State University

Presenters: Zhengyuan Zhu, Gyuhyeong Goh, KyuTae Kim, Jae-kwang Kim

Synthetic Data for Local Statistics: A New Approach to Small Area Estimation"

Zhengyuan Zhu, Iowa State University

Abstract. Public-use survey microdata are often released without detailed geographic identifiers to reduce disclosure risk, making localized inference challenging. This talk presents a synthetic-data framework that integrates survey microdata with domain-level summary statistics and parcel-level information to generate synthetic populations representative of small areas. The proposed approach combines calibration, weight sharing, probabilistic integerization based on the cube method, and spatial allocation to construct geographically coherent synthetic populations while preserving known local constraints. The resulting parcel-level synthetic data support estimation at flexible geographic scales and provide a practical framework for downscaling public-use survey data to local domains. Applications using the American Community Survey and administrative data from Fairfax County, Virginia, demonstrate the method's accuracy and practical value for localized statistical inference.

Weighted Bayesian bootstrap for Bayesian inference under complex sampling

Gyuhyeong Goh, Kyungpook National University

Abstract. Bayesian likelihood-based inference can produce invalid results under informative sampling designs. Existing approximate Bayesian methods for complex survey data typically rely on summary statistics and require the computation of design-based variance estimators. Although the recently

proposed Survey-adjusted Weighted Likelihood Bootstrap (S-WLB) offers a computationally efficient alternative, it may suffer from finite-sample bias and does not incorporate prior information. In this paper, we propose a survey-adjusted weighted Bayesian bootstrap method for Bayesian inference under complex sampling designs. The proposed approach incorporates survey weights and prior information while yielding design-unbiased estimators for the mean and variance of the estimating equations. Simulation studies demonstrate favorable finite-sample performance relative to existing methods, and a real data analysis illustrates the practical applicability of the proposed method.

Predictive Inference for Panel Survey Data under an Autoregressive Model with an Application to the NRI Rangeland On-Site Survey

Kyutae Kim, Iowa State University

Abstract. We propose a predictive inference approach in panel survey data, motivated by the National Resources Inventory (NRI) On-Site Grazingland Project. The finite population means are treated as temporally correlated random quantities, and we develop a two-level model combining (i) a unit-level AR(1) model that captures the time evolution of individual units and the dependence structure induced by revisited panels, and (ii) an area-level AR(1) structure model describing the time evolution of the population means. For parameter estimation, standard Maximum Likelihood Estimation (MLE) targets parameter consistency and does not necessarily yield the best predictor in finite samples. To address this in the second level, we propose a Minimum Predictive Loss (MPL) and Loss-Likelihood Bootstrap (LLB) procedure that treats the autocorrelation parameter as a tuning parameter, selected by minimizing a rolling-origin cross-validated predictive loss. the LLB step further propagates tuning uncertainty into the final predictor. We also derive plug-in MSPE estimators for the resulting predictors. Simulation results show that the proposed MPL/LLB approach achieves smaller mean squared prediction error than standard likelihood-based estimators. We also present the results applying our approach to the NRI Rangeland On-Site Survey data, for predicting population means and quantifying predictive uncertainty.

Bregman projection for calibration estimation in survey sampling

Jae-kwang Kim, Iowa State University

Abstract. Calibration weighting is a fundamental technique in survey sampling and missing data analysis for incorporating auxiliary information and improving estimation efficiency. Classical methods are typically formulated through distance functions on weight ratios relative to design weights. We develop a unified framework for calibration estimation based on Bregman divergence defined directly on the weight vector. Calibration estimators obtained from Bregman divergence admit a dual representation depending only on the dimension of the auxiliary variables, interpretable as a Bregman projection onto the calibration constraint set. This geometric structure yields a general asymptotic expansion showing that calibration estimators are equivalent to debiased regression estimators whose regression coefficient depends on the Bregman generator. This unifies classical methods such as quadratic calibration and exponential tilting, and reveals how the divergence choice influences efficiency. Simulation studies and a real data application illustrate the practical advantages of the proposed approach.

Subsampling and Model Evaluation for Complex Data (IS-24)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 301

Chair: Chenlu Shi

Proposer: Haiying Wang and Chenlu Shi, University of Connecticut, New Jersey Institute of Technology

Presenters: Lin Wang, Dingyi Wang, Jing Wang, Huimin Cheng

Crossed Component Designs for Fast Active-Variable Gaussian Process Emulation in Stochastic Computer Experiments

Lin Wang, Purdue University

Abstract. High-dimensional stochastic computer experiments require both active-variable identification and surrogate modelling of the latent mean response. Standard screening designs provide direct finite-effect information but are poorly structured for Gaussian process emulation, whereas space-filling designs support prediction but give only

indirect screening information and lead to dense covariance matrices. We propose crossed component designs, which partition the variables into groups and evaluate the simulator on the Cartesian product of low-dimensional component designs. The product structure supports exact one-factor screening contrasts and, after active-variable selection, yields a Kronecker covariance matrix for group-separable Gaussian process regression. Likelihood evaluation and prediction can therefore be performed by tensor linear algebra rather than dense Cholesky factorization. Component designs may be Morris, maximum one-factor-at-a-time, Latin hypercube, or hybrid screening–space-filling designs. An auxiliary contrast-variance estimator provides a practical noise estimate for stochastic simulators. Numerical experiments in 60 dimensions show that hybrid crossed designs improve the prediction–screening tradeoff and offer a computationally efficient route to active-variable Gaussian process emulation.

Maximum-Variance-Reduction Stratification for Improved Subsampling

Dingyi Wang, Chinese Academy of Sciences, University of Connecticut

Abstract. Subsampling is a widely used and effective approach for addressing the computational challenges posed by massive datasets. Substantial progress has been made in developing non-uniform, probability-based subsampling schemes that prioritize more informative observations. We propose a novel stratification mechanism that can be combined with existing subsampling designs to further improve estimation efficiency. We establish the estimator's asymptotic normality and quantify the resulting efficiency gains, which enables a principled procedure for selecting stratification variables and interval boundaries that target reductions in asymptotic variance. The resulting algorithm, Maximum-Variance-Reduction Stratification (MVRS), achieves significant improvements in estimation efficiency while incurring only linear additional computational cost. MVRS is applicable to both non-uniform and uniform subsampling methods. Experiments on simulated and real datasets confirm that MVRS markedly reduces estimator variance and improves accuracy compared with existing subsampling methods.

Regression-Calibrated Lasso: Measurement error correction in high-dimensional sparse regression

Jing Wang, University of Pennsylvania

Abstract. The Least Absolute Shrinkage and Selection Operator (Lasso) is a popular tool for performing parameter estimation and variable selection in high-dimensional settings. However, measurement error can preclude Lasso consistency. High-dimensional data with measurement error can be found in nutritional metabolomics, genomics, neuroimaging, and electronic health records studies. We propose a correction method for the effects of measurement error on the Lasso for linear and logistic regression. We retain the favorable convexity of the Lasso and do not require long computation times. We evaluate the efficacy of this method with theory and simulations, both when measurement error and predictor covariance matrices are known, and when they are unknown but can be estimated with replicate data. Convergence rates for our method's estimation error and variable selection consistencies are derived under sub-Gaussian assumptions The estimator's performance is empirically studied with simulations for both linear and logistic regression. Our method performs more reliably than the naive Lasso in the presence of measurement error. We conclude by applying our method to nutritional metabolomics data.

Graphon Cross-Validation: Assessing Models on Network Data

Huimin Cheng, Boston University

Abstract. Graphon models have emerged as powerful tools for modeling complex network structures by capturing connection probabilities among nodes. A key challenge in their application lies in accurately characterizing the graphon function, particularly with respect to parameters that govern its smoothness, which significantly impact the estimation accuracy. In this article, we propose a novel graphon cross- validation method for selecting tuning parameters and estimation approaches. Our method is both theoretically sound and computationally efficient. We show that our proposed cross-validation score is asymptotically parallel to the estimation error, and the selected model asymptotically converges to the optimal model. Through extensive simulations and real-world applications, we demonstrate that our method consistently delivers superior computational efficiency

and accuracy.

StatsUpAI Highlights New and Noteworthy Research at the Intersection of Statistics and Artificial Intelligence (IS-25)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 205

Chair: Jackson Lautier

Proposer: Jackson Lautier, Bentley University

Presenters: Elynn Chen, Qiao Liu, Ying Zhou, Rong Ma

Transfer Q-Learning

Elynn Chen, New York University

Abstract. In dynamic decision-making scenarios across business, healthcare, and education, leveraging data from diverse populations can significantly enhance reinforcement learning (RL) performance for specific target populations, especially when target samples are limited. We develop comprehensive frameworks for transfer learning in RL, addressing both stationary Markov decision processes (MDPs) with iterative Q-learning and non-stationary finite-horizon MDPs with backward inductive learning. For stationary MDPs, we propose an iterative Q-learning algorithm with knowledge transfer, establishing theoretical justifications through faster convergence rates under similarity assumptions. For time-inhomogeneous finite-horizon MDPs, we introduce two key innovations: (1) a novel "reweighted targeting procedure" that enables vertical information-cascading along multiple temporal steps, and (2) transfer deep Q-learning that leverages neural networks as function approximators. We demonstrate that while naive sample pooling strategies may succeed in regression settings, they fail in MDPs, necessitating our more sophisticated approach. We establish theoretical guarantees for both settings, revealing the relationship between statistical performance and MDP task discrepancy. Our analysis illuminates how source and target sample sizes impact transfer effectiveness. The framework accommodates both transferable and non-transferable transition density ratios while assuming reward function transferability. Our analytical techniques have broader implications, extending to supervised transfer learning with neural networks and domain shift scenarios. Empirical evidence from both synthetic and real datasets validates our theoretical results, demonstrating significant improvements over single-task learning rates and highlighting the practical value of strategically constructed transferable RL samples in both stationary and non-stationary contexts. Related papers: Deep Transfer Q-Learning for Offline Non-Stationary Reinforcement Learning: https://arxiv.org/abs/2501.04870 Transfer Q-Learning: https://arxiv.org/abs/2202.04709 Data-Driven Knowledge Transfer in Batch Q-Learning: https://arxiv.org/abs/2404.15209 Transition Transfer Q-Learning for Composite Markov Decision Processes: https://arxiv.org/abs/2502.00534

AI-powered Bayesian Generative Modeling for Statistical Inference

Qiao Liu, Yale University

Abstract. Modern statistical learning increasingly requires methods that can handle complex, high-dimensional data while providing interpretable inference and principled uncertainty quantification. In this talk, I will present a Bayesian generative modeling (BGM) framework for statistical inference. I will first introduce a causal inference approach via Bayesian generative modeling for estimating causal effects from observational data with high-dimensional covariates. By learning low-dimensional latent representations of confounding structure and performing Bayesian latent-variable inference, the model enables causal effect estimation with uncertainty quantification, including individualized and average treatment effects. I will then move to a more general setting that learns the joint distribution of complex data and supports arbitrary conditional inference after training via BGM. This "train once, infer anywhere" paradigm allows a single fitted model to address multiple tasks, such as prediction, missing-data imputation, and multimodal inference, without retraining for each conditioning structure. Together, these models illustrate how Bayesian generative modeling (BGM) can unify modern AI and statistics by combining flexible neural architectures with statistically principled inference.

Learning from the Unseen: Offline Reinforcement Learning with Hidden Actions

Ying Zhou, University of Connecticut

Abstract. Offline reinforcement learning typically assumes that actions in the dataset are observed without error. In many applications, however, the true actions may be unobserved and only noisy

proxies are available, leading to bias in standard off-policy evaluation and potentially misleading conclusions. We study off-policy evaluation in infinite-horizon discounted Markov decision processes with hidden actions. By leveraging the next-state variable as a natural proxy for the unobserved action, we establish identification of the policy value and propose an influence function-based estimator, LURE (**L**earning from the **U**nseen: **R**obust **E**stimator). The LURE estimator is multiply robust, remaining consistent under several combinations of correctly specified nuisance components, and is asymptotically normal, enabling valid inference. To our knowledge, this is the first work on offline reinforcement learning with hidden actions. Simulations and a sepsis management application using the MIMIC-III database show that LURE substantially reduces bias compared to baseline methods.

Kernel Spectral Joint Embeddings for High-Dimensional Noisy Datasets Using Duo-Landmark Integral Operators

Rong Ma, Harvard University

Abstract. Integrative analysis of multiple heterogeneous datasets has arised in many research fields. Existing approaches oftentimes suffer from limited power in capturing nonlinear structures, insufficient account of noisiness and effects of high-dimensionality, lack of adaptivity to signals and sample sizes imbalance, and their results are sometimes difficult to interpret. To address these limitations, we propose a kernel spectral method that achieves joint embeddings of two independently observed high-dimensional noisy datasets. The proposed method automatically captures and leverages shared low-dimensional structures across datasets to enhance embedding quality. The obtained low-dimensional embeddings can be used for downstream tasks such as simultaneous clustering, data visualization, and denoising. The proposed method is justified by rigorous theoretical analysis, which guarantees its consistency in capturing the signal structures, and provides a geometric interpretation of the embeddings. Under a joint manifolds model framework, we establish the convergence of the embeddings to the eigenfunctions of some natural integral operators. These operators, referred to as duo-landmark integral operators, are defined by the convolutional kernel maps of some reproducing kernel Hilbert spaces (RKHSs). These RKHSs capture the underlying, shared low-dimensional nonlinear signal structures between the two datasets. Our numerical experiments and analyses of two pairs of single-cell omics datasets demonstrate the empirical advantages of the proposed method over existing methods in both embeddings and several downstream tasks. Supplementary materials for this article are available online, including a standardized description of the materials available for reproducing the work.

Statistics and AI for Science and Society (IS-29)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 110

Chair: Donghui Yan

Proposer: Donghui Yan, UMass Dartmouth

Presenters: Hanna Yan, Rachel Yang, Mabelle Liu, Alex Gan, Lawrence Du

VCEP Aware AI Agent for Epilepsy Sodium Channel Variant Classification

Hanna Yunfei Yan, Basis Independent Silicon Valley

Abstract. Variant Curation Expert Panels (VCEPs) within ClinGen evaluate the pathogenicity of genetic variants by integrating evidence from population databases, clinical cases, computational predictors, and functional studies. This process demands extensive manual literature review and multi-tiered approval, making it poorly suited to the rapidly expanding volume of genomic data. We present an AI agent that automates variant classification for voltage-gated sodium channel genes while preserving the accuracy of human expert panels. We focused initially on PVS1, the strongest evidence category in the ACMG/AMP framework, which designates very strong pathogenic evidence for loss-of-function (LOF) variants. Assessing PVS1 requires navigating a branched decision tree that incorporates nonsense-mediated decay (NMD) prediction and additional factors contributing to pathogenicity. AutoPVS1 (Xiang et al., 2020) automates this decision process. but operates exclusively on VCF-format genomic coordinates, whereas VCEPs and clinical databases use transcript-based HGVS nomenclature. For a given genomic coordinate, AutoPVS1 selects a single default transcript per gene. In genes such as SCN8A, which have multiple transcripts with differing exon structures, this default selection can misclassify a LOF variant as non-LOF and therefore non-PVS1. For example, a canonical splice-acceptor variant on NM_014191.4 meets

PVS1 criteria, yet AutoPVS1 reports it under NM_001330260.2 as a deep intronic, non-PVS1 variant — despite both annotations referring to the same nucleotide substitution and genomic position. Our agent addresses this limitation through a wrapper that accepts HGVS input, converts it to VCF format for positional mapping, and then resolves the evaluation against the user-specified transcript, preventing misannotation. The agent also introduces an AI-derived confidence score for biologically uncertain cases, flagging variants for expert review when, for instance, more than 10% of the encoded protein would be truncated. Validation against 63 variants from the ClinGen Epilepsy Sodium Channel VCEP Evidence Repository yielded 96.8% accuracy (61/63). All high -confidence classifications matched VCEP consensus. The two discordant cases involved biologically ambiguous variants where expert panel interpretation had itself diverged from the standard decision tree.; the agent correctly flagged both for review. Future work will extend the agent to REVEL -dependent criteria, including PP3 and BP4 and PM 1-6.

Unified Deep Learning Framework for MRI Image Contrast Translation

Rachel Yang, Scarsdale High School

Abstract. Despite advances in imaging technology in healthcare, several limitations remain in medical imaging and diagnosis. Constraints of physical imaging systems, including patient tolerance, acquisition time, and high costs, can prevent patients from receiving accurate and timely diagnoses. MRI image contrast translation, such as generating T1-weighted images from T2-weighted images or vice versa, offers a potential solution to improve clinical workflow and reduce the need for multiple imaging sequences. In this work, we develop a unified deep learning-based image-to-image translation framework to synthesize T2-weighted images from T1-weighted images and vice versa. The model is based on a U-Net architecture and is trained using a combination of loss functions, including L1, gradient, perceptual, and frequency losses, to preserve both image contrast and structural details. A total of 40 brain MRI datasets, each containing paired T1- and T2-weighted images, were used for training. Through controlled experiments that systematically vary individual components, we evaluate the impact of model design choices and training parameters on image quality, including clarity, structural accuracy, and reconstruction fidelity, and identify optimal model performance. Preliminary results demonstrate strong quantitative performance of the proposed model, achieving a Structural Similarity Index Measure (SSIM) of up to 0.9485 for T1-to-T2 translation and 0.9913 for T2-to-T1 translation, as well as a Peak Signal-to-Noise Ratio (PSNR) of 26.34 dB 27.35 dB, respectively. These findings suggest that the proposed approach can reliably synthesize high-quality MRI contrasts and has the potential to reduce scan time, improve patient comfort, and enhance the efficiency of clinical MRI workflows.

Intentions vs. Actions: Effects of Personal and Social Drivers of Youth Activism Regarding Banned Ethnic Studies Books Among High School Students

Mabelle Liu, Marriotts Ridge High School

Abstract. Ethnic Studies (ES) courses and books are empirically advantageous for both traditionally marginalized students and their White counterparts. Since the 1960s, high school students' advocacy efforts have been crucial for the inception and preservation of ES pedagogy. However, the underlying motivation behind youth activism remains under-explored in existing literature during a time where contemporary policies increasingly restrict ES by banning books essential to its curricula. This mixed-method study examines high school students' perceptions on ES book bans and youth activism to understand what drives their behavioral intention for and actual involvement in advocacy. Analyses of 192 survey responses show that while students' personal perceptions on ES, censorship, and activism predict their behavioral intention, the perceptions of attitudes and behaviors of parents, peers, and teachers predict students' actual involvement in advocacy. These findings suggest that effective mobilization against book bans requires stronger community commitment to encouraging student action.

Uncovering Hidden Housing Insecurity: A Monte Carlo and Bayesian Framework for Enhancing PIT Homelessness Counts

Alex Gan, Mclean High School

Abstract. In the field of social studies, Point-in-Time (PIT) counts have been used as a standard method for measuring homelessness across regions. However, this method captures a static, single-night snapshot of the local homelessness situation. Consequently, it may miss individuals who are experiencing hidden, temporary, or intermittent housing insecurity. To

address potential PIT undercounting issues in Fairfax County, Virginia, this research analyzes an array of county-level economic indicators from Federal Reserve Economic Data (FRED), including housing inventory, mortgage rates, unemployment, and housing prices, as well as building permits. It attempts to identify the structural and economic pressures that drive up localized housing insecurity. Based on economic housing-pressure indicators, this study develops a Bayesian evidence-synthesis model to refine conservative baseline estimates from the PIT data. A Monte Carlo simulation is then used to calculate the range of uncertainty for the new homelessness estimate. The official PIT count in Fairfax County for 2025 recorded 1,322 individuals, likely understating the full extent of housing insecurity. On average, the model suggests the adjusted estimate was approximately 131% higher than the PIT count. Although these figures are not meant to replace official counts, they demonstrate the power of statistical modeling to uncover hidden housing issues among the most vulnerable groups in Fairfax County. The proposed methodology also provides a more effective guide for local policy planning. This approach highlights the immediate need for local governments to integrate PIT counts with shelter usage, eviction records, and economic indicators to create a more frequent and accurate database.

Statistical Analysis of Student Success in Towson University's Intermediate Algebra Course

Lawrence Du, Gilman School

Abstract. Introductory mathematics courses often determine students' long-term academic success, making it essential to understand the factors that shape early success. This study analyzes 1,839 first-attempt enrollments in MATH 102 from Fall 2021 to Fall 2024 to examine how demographic characteristics, academic preparation, and enrollment pathways influence student outcomes. Using logistic regressions, a fundamental machine-learning/AI model, and Chi-squared tests, the results reveal substantial racial disparities and lower success rates for First-generation and Pell-eligible students, while also showing that High school GPA was strongly correlated to success in MATH 102. On the contrary, minimum ALEKS placement scores were not correlated with course outcomes. Analysis of enrollment pathways showed that students who completed recommended developmental prerequisites, such as MATH 95 or ORIE 101, achieved results similar to those of direct-entry students. In contrast, students who went against department-recommended pathways had severely reduced success rates. Course performance also had downstream effects: students who earned a DFW were significantly more likely to change majors. These findings highlight the need for strengthened advising and greater support for vulnerable student groups to improve retention and success in foundational mathematics.

Innovations in Multivariate Methods for Biomedical Research (IS-34)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 305

Chair: Victor Hugo Lachos Davila

Proposer: Jungwun Lee, Boston University School of Public Health

Presenters: Jungwun Lee, Benjamin Stockton, Sreeram Anantharaman, Wei Jin

A latent class profile selection model for non-ignorable missing values in longitudinal data: application to physical health and function trajectories in adults with cancer

Jungwun Lee, Boston University School of Public Health

Abstract. This paper proposes a novel latent trajectory method for dealing with non-responses that are possibly non-ignorable when identifying latent trajectories of multivariate categorical outcomes. Specifically, the proposed trajectory model describes the joint distribution of multivariate longitudinal outcomes and their missing patterns using two types of categorical latent variables. A latent trajectory variable summarizes the longitudinal response patterns of multivariate outcomes, and a latent missingness variable summarizes the missingness patterns of multivariate outcomes. In this way, subjects with the same latent trajectory and missingness memberships will share common response and non-response patterns, whereas subjects with different latent memberships are heterogeneous. We employ the Expectation-Maximization algorithm to obtain maximum-likelihood estimates. We demonstrate the novelty of the proposed model via simulation studies and by analyzing the YUCAN data set, a prospective cohort study on adult patients diagnosed with cancer between 2019 and 2022.

Clarifying the role of placebo response classification in the analysis of the Sequential Parallel Comparison Design

Benjamin Stockton, Division of Biostatistics, Department of Population Health, NYU Grossman School of Medicine

Abstract. Sequential parallel comparison design (SPCD) clinical trials aim to adjust active treatment effect estimates for placebo response to minimize the impact of placebo responders on the estimates. This is potentially accomplished using a two stage design by measuring treatment effects among all participants during the first stage, then classifying some placebo arm participants as placebo non-responders who will be re-randomized in the second stage. In this paper, we use causal inference tools to clarify under what assumptions treatment effects can be identified in SPCD trials and what effects the conventional estimators target at each stage of the SPCD trial. We further illustrate the highly influential impact of placebo response misclassification on the second stage estimate. We conclude that the conventional SPCD estimators do not target meaningful treatment effects.

Predicting HIV-1 viral load for monitoring treatment failure considering gaps in measurement

Sreeram Anantharaman, Brown University

Abstract. Viral load (VL) monitoring is essential for identifying virologic failure in HIV care, but routine program data often contain incomplete records because patients frequently miss or delay scheduled VL tests. Most existing research focuses on patients with complete VL follow-up and simply excludes those with partial or irregular records, potentially biasing estimates of virologic failure and limiting the applicability of findings to real-world settings. To address this gap, we develop a joint modeling framework that incorporates all patients after ART initiation regardless of completeness of follow-up by linking VL failure, defined as VL >=1000 copies/mL, with the timing of VL measurements. VL failure is modeled via logistic regression, and the measurement-time process is modeled using a Random Survival Forest that flexibly captures nonlinear covariate effects and heterogeneity in testing patterns. The two components are combined through an exponential-tilt pattern-mixture structure, enabling delayed or missed VL tests to inform estimation and prediction. We illustrate the methods using data from a large HIV care program in Kenya, showing how the approach enables individual-level prediction of VL failure among patients with irregular monitoring, while also allowing population- and clinic-level estimates that reflect incomplete follow-up.

Learning Optimal Dynamic Treatment Regimes Under Unmeasured Confounding

Wei Jin, Department of Biostatistics, Boston University School of Public Health

Abstract. Data-driven personalized decision-making has become increasingly important in many scientific fields. Existing methods often rely on the assumption of no unmeasured confounding for identifying the optimal dynamic treatment regimes (DTRs). However, this assumption is often violated in practice, especially in observational studies. While techniques like instrumental variables or proxy variables can help address unmeasured confounding, such additional data sources are not always available. In this talk, we propose a novel Bayesian approach for learning optimal DTRs with continuous treatments under unmeasured confounding. For causal identification, we propose a Bayesian dynamic causal model that achieves unique identification under certain mild distributional assumptions, without requiring additional data sources. For policy optimization, we develop a practical algorithm that robustly learns the optimal DTRs by identifying a conservative policy. Through simulations and an application to a large-scale kidney transplantation dataset, we demonstrate the proposed method's identifiability, utility, and robustness, highlighting its value in advancing precision medicine.

Robust Inference: Model Selection, Causal Analysis, and Quantile Regression (IS-36)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 306

Chair: Zhen Chen

Proposer: Ruijin Lu, Washington University in St. Louis School of Medicine

Presenters: Jongwoo Choi, Zhejia Dong, Sean O'Hagan

Robust model selection using likelihood as data

Jongwoo Choi, University of Connecticut

Abstract. Model selection is a central task in statistics, but standard methods are not robust in misspecified settings where the true data-generating process (DGP) is not in the set of candidate models. The key limitation is that existing methods—including information criteria and Bayesian posteriors—do not quantify uncertainty about how well each candidate model approximates the true DGP. In this paper, we introduce a novel approach to model selection based on modeling the likelihood values themselves. Specifically, given$K$candidate models and$n$observations, we view the$n \times K$matrix of negative log-likelihood values as a random data matrix and observe that the expectation of each row is equal to the vector of Kullback–Leibler divergences between the$K$models and the true DGP, up to an additive constant. We use a multivariate normal model to estimate and quantify uncertainty in this expectation, providing calibrated inferences for robust model selection under misspecification. The procedure is easy to compute, interpretable, and comes with theoretical guarantees, including consistency.

Design and analysis for valid causal inference with network-dependent data

Zhejia Dong, Brown University

Abstract. Matching is widely used to mimic randomized experiments by forming matched sets in which treated and control units differ only randomly with respect to important confounding variables. However, when the study population consists of interconnected units from a single network or a small number of networks (e.g., social or proximity networks), matching solely on confounding variables may produce matched units that are not randomly different with respect to their network distance, but instead are more likely to be closely connected after matching. Such increased network closeness within matched sets may induce spurious associations between treatment and outcome, when both variables exhibit shared autocorrelation patterns on the network. To reduce spurious associations within matched sets while preserving the validity of within-matched-set causal comparisons, we propose a new matching method that matches units with similar covariates while reducing within-matched-set dependence by imposing additional constraints on network proximity. Furthermore, at the analysis stage, to account for residual dependence across matched sets, we propose a valid randomization inference procedure for testing the sharp null hypothesis of no causal effect that accommodates across-matched-set dependence without explicit assumptions on the underlying dependence structure. We demonstrate the validity and utility of the proposed methods through simulation studies and an application to real-world social network data.

Generative Regression with IQ-BART

Sean O'Hagan, University of Chicago

Abstract. Implicit Quantile Bayesian Additive Regression Trees (IQ-BART) posits a non-parametric Bayesian model on the conditional quantile function, acting as a model over a conditional model for$Y$given$\mathbf{X}$. Using the fact that the location parameterμin aτ-asymmetric Laplace distribution corresponds to theτ^{th}quantile, we build a check-loss likelihood targetingμas the parameter of interest. We equip the check-loss likelihood parameterized by$\mu = f(\mathbf{X}, \tau)$with a constrained BART prior on$f(\cdot)$, allowing the conditional quantile function to vary both in$\mathbf{X}$andτ. The posterior distribution over$f$can be then distilled for estimation of the entire conditional quantile function as well as for assessing uncertainty through the variation of posterior draws. Simulation-based predictive inference is immediately available through inverse transform sampling using the learned quantile function. The sum-of-trees structure over the conditional quantile function enables flexible non-parametric regression with theoretical guarantees. We demonstrate the power of IQ-BART on time series forecasting datasets where IQ-BART can capture multimodality in predictive distributions that might be otherwise missed using traditional parametric approaches.

New perspectives on Bayesian computations: approximation, sampling, and diffusion models (IS-39)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 206

Chair: Nianqiao Phyllis Ju

Proposer: Nianqiao Phyllis Ju, Dartmouth College

Presenters: Yves Atchade, Guanyang Wang, Sifan Liu, Diana Cai

One-step Laplace Approximation for Bayesian Variable Selection

Yves Atchade, Boston University

Abstract. Feature selection in high-dimensional settings is a central challenge in modern science

and decision-making. Existing methods with strong statistical guarantees are often computationally intractable at scale, limiting their practical utility. To bridge this gap, we propose OLAP (One-step LAplace approximation), a novel method built on Le Cam's one-step procedure that is designed to preserve statistical guarantees while substantially reducing computational cost. We establish that, under standard high-dimensional assumptions, OLAP achieves consistent variable selection. The method further delivers a posterior distribution that can be efficiently explored in polynomial time using a Gibbs sampling algorithm. The practical performance of OLAP is demonstrated through applications to logistic and Poisson regression on both simulated and real data.

Weak Diffusion Priors Can Still Achieve Strong Inverse-Problem Performance

Guanyang Wang, Rutgers University

Abstract. Can a diffusion model trained on bedrooms recover human faces? Diffusion models are widely used as priors for inverse problems, but standard approaches usually assume a high-fidelity model trained on data that closely match the unknown signal. In practice, one often must use a mismatched or low-fidelity diffusion prior. Surprisingly, these weak priors often perform nearly as well as full-strength, in-domain baselines. We study when and why inverse solvers are robust to weak diffusion priors. Through extensive experiments, we find that weak priors succeed when measurements are highly informative (e.g., many observed pixels), and we identify regimes where they fail. To explain this behavior, we combine Bayesian-consistency theory with local-correlation analysis: the theory gives conditions under which high-dimensional measurements make the posterior concentrate near the true signal, while the correlation analysis shows that weak and stronger natural-image priors can share similar local spatial structure. These results provide a principled justification on when weak diffusion priors can be used reliably. Code is available at https://github.com/jjia131/weak-diffusion-priors-inverse-problem .

Rotated Mean-Field Variational Inference and Iterative Gaussianization

Sifan Liu, Duke University

Abstract. Mean-field variational inference (MFVI) approximates a target distribution with a product distribution in the standard coordinate system, offering a scalable approach to Bayesian inference but often severely underestimating uncertainty due to neglected dependence. We show that MFVI can be greatly improved when performed along carefully chosen principal component axes rather than the standard coordinates. The principal components are obtained from a cross-covariance matrix of the target's score function and identify orthogonal directions that capture the dominant discrepancies between the target distribution and a Gaussian reference. Performing MFVI in a rotated system defines a rotation followed by a coordinatewise transformation that moves the target closer to Gaussian. Iterating this procedure yields a sequence of transformations that progressively Gaussianize the target. The resulting algorithm provides a computationally efficient construction of normalizing flows, requiring only MFVI sub-problems and avoiding large-scale optimization. In posterior sampling tasks, we demonstrate that the proposed method greatly outperforms standard MFVI while achieving accuracy comparable to normalizing flows at a much lower computational cost.

Fisher meets Feynman: score-based variational inference with a product of experts

Diana Cai, Flatiron Institute

Abstract. We introduce a highly expressive yet distinctly tractable family for black-box variational inference (BBVI). Each member of this family is a weighted product of experts (PoE), and each weighted expert in the product is proportional to a multivariate-distribution. These products of experts can model distributions with skew, heavy tails, and multiple modes, but to use them for BBVI, we must be able to sample from their densities. We show how to do this by reformulating these products of experts as latent variable models with auxiliary Dirichlet random variables. These Dirichlet variables emerge from a Feynman identity, originally developed for loop integrals in quantum field theory, that expresses the product of multiple fractions (or in our case, -distributions) as an integral over the simplex. We leverage this simplicial latent space to draw weighted samples from these products of experts—samples which BBVI then uses to find the PoE that best approximates a target density. Given a collection of experts, we derive an iterative procedure to optimize the exponents that determine their geometric weighting in the PoE. At each iteration, this procedure minimizes a regularized Fisher divergence to match the scores of the variational and target densities at a batch

of samples drawn from the current approximation. This minimization reduces to a convex quadratic program, and we prove under general conditions that these updates converge exponentially fast to a near-optimal weighting of experts. We conclude by evaluating this approach on a variety of synthetic and real-world target distributions.

Statistics in Neuroimaging (IS-49)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 202

Chair: Zan Li

Proposer: Spencer Wadsworth, University of Connecticut

Presenters: Spencer Wadsworth, Xiaomeng Ju, Seonjoo Lee

Bayesian Sparsity Modeling of Shared Neural Response Naturalistic fMRI Data

Spencer Wadsworth, University of Connecticut

Abstract. Detecting shared neural activity from functional magnetic resonance imaging (fMRI) across individuals exposed to the same stimulus provides insight into synchronous brain responses and functional organization. Intersubject correlation (ISC) is the standard approach for identifying shared responses under naturalistic stimuli, but it relies on heavy data summarization, involves thousands of tests, and does not directly estimate a shared neural response (SNR) function. We propose a Bayesian model-based alternative that simultaneously identifies spatial regions of shared activity and estimates the SNR function. The approach incorporates a sparse Gaussian process prior to estimate the SNR along with a horseshoe-inspired prior for voxel-level activation with a spatial component on the shrinkage. In a simulation study and analyses on two real fMRI datasets, we demonstrate improved activation detection and response estimation relative to ISC.

Bayesian Modeling for Multilevel Functional Data with Applications to Neuroimaging Analysis

Xiaomeng Ju, NYU Grossman School of Medicine

Abstract. In multi-condition experiments, brain activity is recorded as subjects engage in various tasks or respond to different stimuli. The resulting signals are often transformed into time–frequency representations, which can be viewed as two-way functional data with experimental conditions nested within subjects. Motivated by the analysis of evoked electroencephalogram (EEG) signals, we develop Bayesian mixed-effects models for such time–frequency representations. The proposed models jointly account for the data's multilevel structure, functional nature, and subject-level covariates by incorporating covariate-dependent fixed effects and multilevel random effects. To enhance interpretability and parsimony, we introduce a novel decomposition of the fixed effects that yields marginally interpretable time and frequency patterns, together with a sparsity-inducing prior for rank selection. The proposed approach is evaluated through simulations and applied to EEG data examining the effects of alcoholism on cognitive processing in response to visual stimuli. The talk will also discuss an extension of the framework to the analysis of dynamic functional connectivity.

Longitudinal Canonical Correlation Analysis

Seonjoo Lee, Columbia University

Abstract. We propose a canonical correlation analysis for two longitudinal variables that are possibly sampled at different time resolutions with irregular grids. We modeled trajectories of the multivariate variables using random effects and found the most correlated sets of linear combinations in the latent space. Our numerical simulations showed that the longitudinal canonical correlation analysis (LCCA) effectively recovers underlying correlation patterns between two high-dimensional longitudinal data sets. We applied the proposed LCCA to data from the Alzheimer's Disease Neuroimaging Initiative and identified the longitudinal profiles of morphological brain changes and amyloid cumulation. We will discuss recent extension of the LCCA for mixed type data.

Unpacking Complex Interventions: Causal Inference for Interference and Mediation (IS-52)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 109

Chair: Ashley Buchanan

Proposer: Youjun Li, College of Pharmacy, University of Rhode Island

Presenters: Ivvone Zhou, Jiaqi Tong, Ke Zhang.

Direct and Spillover Effects of Providers' Intended Participation in the STOP-HPV Cluster-randomized Trial

Ivvone Zhou, Perelman School of Medicine, University of Pennsylvania

Abstract. Cluster-randomized trials are often conducted to evaluate interventions targeting staff within an organization such as a hospital or primary care practice. Such interventions are often subject to incomplete staff engagement, leading policymakers to question how staff participation limits or amplifies intervention effectiveness. Using the STOP-HPV study, a cluster-randomized trial conducted to evaluate a sequential, multimodal intervention to reduce missed opportunities for HPV vaccination, we propose inverse-probability weighted estimators of the joint effect and decomposed effects of the practice-level intervention and provider-level intent to participate on the visit-level outcome of missed vaccination opportunity. We index estimands by the joint exposure of the intervention, the visit provider participation, and other-provider participation and marginalize over other-provider participation. We decompose the joint effect into the effect of participation if in the intervention arm (the direct participation effect) and the effect of the intervention if not participating (the indirect spillover effect). We then outline the identifiability assumptions, describe the estimation procedure, and report our findings. This work provides a framework for analyzing spillover effects and their magnitude relative to direct effects in cluster-randomized trials when an indicator of uptake is collected before randomization.

Causal mediation in cluster-randomized trials with multiple mediators: spillover-aware decomposition, identification, and semiparametric efficient inference

Jiaqi Tong, Yale University

Abstract. Causal mediation analysis in cluster-randomized trials (CRTs) is complicated by the presence of multiple mediators, intracluster correlation, and within-cluster interference. Existing mediation methods often fall short in accommodating these features simultaneously, and semiparametric efficient estimators that fully address them remain unavailable. We develop a unified framework that defines a class of mediation effect estimands, including exit indirect effects, exit spillover mediation effects, and their interaction effects, to investigate causal mechanisms in CRTs with an arbitrary number of mediators under an unknown causal structure. We introduce a set of interpretable causal assumptions for point identification of each estimand. For optimal inference, we first derive the efficient influence functions for the proposed estimands and construct corresponding one-step and debiased machine learning estimators. In particular, to flexibly model the joint mediator density, we employ an elliptical copula marginal regression model that combines a nonparametric marginal regression with an interpretable association structure. We assess the finite-sample performance of the proposed estimators through simulation studies and illustrate the methodology by reanalyzing the PPACT CRT data with three causally unordered mediators.

Assessing Direct and Spillover Effects of HIV Testing on HIV Incidence in Rural KwaZulu-Natal, South Africa

Ke Zhang, University of Rhode Island

Abstract. Although there are effective strategies to control the HIV epidemic, it remains a significant individual and public health challenge in South Africa. HIV testing is the gateway to HIV treatment in those who have acquired HIV and HIV prevention in those who tested HIV negative. Existing studies have suggested that HIV testing has a significant effect in reducing HIV incidence. However, these studies have not fully assessed spillover effects, the effects of one's HIV testing on HIV incidence among unexposed others. Assessing spillover can provide a more complete understanding of the impact of HIV testing. The data we used is from ANRS 12249 treatment as prevention (TasP) trial, conducted in a rural region of South Africa from March 2012 to July 2016. We grouped participants by homesteads and assumed partial interference limited to the homestead, estimated both the direct (i.e., the intervention effect under exposure versus no exposure while holding other factors constant) and spillover effects of altering the proportion of HIV testing in the homestead on subsequent HIV incidence. Estimation was carried out with a marginal structured model fit with inverse probability weights. The results suggested that, on average, (1) Risk of HIV infection decreased by 34.5% (95% confidence interval: [19.1%, 49.4%]) under HIV testing versus no testing, with HIV testing for others in the homestead remained constant at 100% (i.e., direct effect); (2) No statistically significant spillover effect detected comparing two different proportions of exposed homestead members. Further research is needed to understand the under-

lying mechanisms.

Reliable Statistical and AI Methods for Clinical Decision-Making Using EHR and Medical Data (IS-58)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 111

Chair: JooChul Lee

Proposer: JooChul Lee, Auburn University

Presenters: JooChul Lee, Weidong Ma, Xingyan Li, Hong Xiong

Evidence-grounded Clinical Question Answering with Probabilistic Region Selection

JooChul Lee, Auburn University

Abstract. Clinical visual question answering (VQA) on chest X-rays should not only produce correct answers but also identify the image evidence supporting each prediction. However, existing grounded medical VQA systems are typically formulated as region localization or attention-weighting problems, rather than as evidence inference tasks. Consequently, they do not explicitly model whether sufficient evidence is present to support a prediction, particularly in disease-absent cases. In this work, we propose an evidence-grounded clinical question answering framework that formulates visual grounding as an evidence inference problem. Instead of normalized attention, the model estimates independent relevance probabilities for candidate regions conditioned on the clinical question. This formulation improves alignment between predicted answers and their supporting evidence while reducing spurious visual grounding, and maintains competitive answer performance. Our results highlight the importance of explicitly modeling and controlling visual evidence, including the absence of evidence, for building reliable and clinically interpretable medical AI systems.

The Breakdown Point and Asymptotic Relative Efficiency of Hybrid Wilcoxon Test under a Mixed Design

Xingyan Li, UT Health Houston

Abstract. Hybrid Wilcoxon test for clinical trials under mixed designs provide a novel statistical approach for testing and estimating the overall location parameter of two distributions arising from both matched and unmatched sub-studies. The method is based on the Wilcoxon rank test and is expected to be robust while maintaining reasonable efficiency compared with parametric tests on the corresponding datasets. The breakdown point, defined as the smallest proportion of contaminated samples that can cause an estimator to take arbitrarily large aberrant values, is used to assess robustness. Asymptotic relative efficiency (ARE), defined as the ratio of efficiencies between two tests when the null and alternative location parameters converge to the same value, is used to compare the performance of two tests under the same sampling distribution. In this presentation, I will derive and show the computation, for each sub-designs and for the mixed design: 1. the theoretical breakdown points of both the estimator and its variance; 2. the theoretical ARE between the t-type test and the corresponding Wilcoxon test. The simulation results under both large and small sample sizes is used to evaluate the accuracy of the theoretical breakdown points, and ARE between the meta t-type test and the hybrid Wilcoxon.

Equivalence Testing for Algorithmic Fairness in Clinical Risk Prediction

Hong Xiong, University of Pennsylvania

Abstract. Fairness assessments in clinical risk prediction often compare subgroup true positive rates (TPRs) and test whether the difference is exactly zero. This can be misleading: TPR differences may reflect subgroup differences in underlying risk rather than model performance, and failure to reject the null hypothesis of equal TPRs does not establish that any true subgroup difference in TPRs is practically negligible. We address both issues by leveraging the adjusted true positive rate (aTPR), which evaluates subgroup performance counterfactually under a common reference risk distribution, and by adapting equivalence-testing logic to a Wald procedure for the aTPR contrast, Two One-Sided Wald Tests (TOSW). TOSW tests the null hypothesis that the subgroup difference in aTPR is at least as large in absolute value as a prespecified equivalence margin. We evaluate TOSW using simulated data as well as a semi-synthetic dataset derived from the Veterans Affairs (VA) electronic health records to assess the fairness of the VA Care Assessment Needs (CAN) algorithm for predicting one-year hospitalization. Simulation results showed that, in larger samples, TOSW had empirical type I error close to the nominal level at the equivalence boundary and power that increased as the magnitude of the true subgroup difference decreased from the equivalence boundary toward

zero, whereas conventional TPR-based testing declared equivalence too often when the target inequity was defined on the aTPR scale. In the semi-synthetic study, under covariate shift that preserved the relationship between predicted risk and observed outcomes, both metrics continued to support equivalence, providing a useful baseline validation of TOSW. Under calibration drift, by contrast, aTPR detected loss of equivalence at smaller drift magnitudes and more consistently across repeated samples than TPR. Overall, aTPR with TOSW provides an interpretable and practically relevant method for fairness assessment in clinical risk prediction.

Modern Principled Learning for Causal Inference and.Decision-Making (IS-77)

Time & Location: May 29, 01:30 PM - 03:10 PM | Room 108

Chair: Qingkai Dong

Proposer: Xingche Guo, University of Connecticut

Presenters: Diptanil Santra, Junhui Yang, Qingkai Dong

Distributional Balancing for Causal Inference: A Unified Framework via Characteristic Function Distance

Diptanil Santra, University of Illinois at Urbana-Champaign

Abstract. Weighting methods are essential tools for estimating causal effects in observational studies, with the goal of balancing pre-treatment covariates across treatment groups. Traditional approaches pursue this objective indirectly, for example, via inverse propensity score weighting or by matching a finite number of covariate moments, and therefore do not guarantee balance of the full joint covariate distributions. Recently, distributional balancing methods have emerged as robust, nonparametric alternatives that directly target alignment of entire covariate distributions, but they lack a unified framework, formal theoretical guarantees, and valid inferential procedures. We introduce a unified framework for nonparametric distributional balancing based on the characteristic function distance (CFD) and show that widely used discrepancy measures, including the maximum mean discrepancy and energy distance, arise as special cases. Our theoretical analysis establishes conditions under which the resulting CFD-based weighting estimator achieves$\sqrt{n}$-consistency. Since the standard bootstrap may fail for this estimator, we propose subsampling as a valid alternative for inference. We further extend our approach to an instrumental variable setting to address potential unmeasured confounding. Finally, we evaluate the performance of our method through simulation studies and a real-world application, where the proposed estimator performs well and exhibits results consistent with our theoretical predictions.

Robust and Adaptive Causal Inference Under Model Uncertainty

Junhui Yang, University of Massachusetts Amherst

Abstract. In observational studies, analysts can often pose multiple plausible causal models for a single dataset. However, each model may rely on untestable assumptions, and it is often unknown which (if any) of them are correctly specified. Since inference based on a single model can be fragile, we develop hypothesis tests that combine evidence across multiple candidate causal models to achieve *causal robustness*: the test remains valid provided at least one candidate model is correct. For each model, we construct a semiparametric, asymptotically linear estimator of its identifying functional and exploit the resulting joint asymptotic normality. This yields a family of sum–product tests that are valid whenever exactly$J$functionals are zero under the null, where$J \geq 1$is fixed and known. We then propose an adaptive test that estimates$J$from the data and uses the corresponding fixed-Jstatistic for testing. We show that the adaptive test asymptotically controls type I error, has power approaching one under fixed alternatives, and attains local power under$n^{-1/2}$alternatives without specialized post-selection adjustments. We further show that inverting the adaptive test yields confidence sets that incorporate both statistical uncertainty and uncertainty about which candidate model(s) are correct. We illustrate our methods using numerical studies and an analysis of the Wisconsin Longitudinal Study.

Preserving Rare Features in Big Data Regression: Balanced Subsampling

Qingkai Dong, University of Connecticut

Abstract. Rare binary covariates, or rare features, pose significant challenges for statistical modeling, including numerical instability and slow convergence. The problem becomes more severe in

large-scale data analysis, where subsampling is often used for computational feasibility but may discard observations carrying rare-feature information. This paper develops subsampling methods for regression models with rare features. We first show that the convergence rate of rare-feature coefficient estimators is determined by the number of observations expressing the corresponding feature, rather than by the full-data sample size. We then establish asymptotic results for general subsample estimators and characterize how subsampling designs affect identifiability and estimation efficiency. Motivated by these results, we propose a rarity-aware extension of classical L-optimal subsampling that preserves rare-feature information. We further introduce a balanced subsampling method that quantifies rarity through a balance score and can be used as either a standalone design or a robust pilot for subsequent optimal subsampling. To incorporate pilot samples while accounting for non-negligible overlap between pilot and second-step samples, we develop a union-sample aggregation procedure and establish its asymptotic normality. Through theoretical analysis, simulations, and real-data applications, we demonstrate that the proposed methods substantially improve estimation stability and efficiency compared with existing approaches.