The 37th New England Statistics Symposium

Program Book

Contents

Parallel Session 1 | 08:30 AM - 10:10 AM, May 22

Recent Advances in Statistical Methods for Neuroimaging Research (IS-2)

Chair: Panpan Zhang

Proposer: Panpan Zhang, Vanderbilt University Medical Center

Room: McHugh 206

Presenters: Guanqun Cao; Simon Vandekar; Suprateek Kundu; Panpan Zhang

Simultaneous Classification and Feature Selection for Complex Functional Data

Guanqun Cao, Michigan State University

Abstract: The opportunity to utilize complex functional data types for conducting classification tasks is emerging with the growing availability of imaging data. However, the tools capable of effectively managing imaging data are limited, let alone those that can further leverage other one-dimensional functional data. Inspired by the extensive data provided by the Alzheimer's Disease Neuroimaging Initiative (ADNI), we introduce a novel classifier tailored for complex functional data. Each observation in this framework may be associated with numerous functional processes, varying in dimensions, such as curves and images. Each predictor is a random element in an infinite dimensional function space, and the number of functional predictors p can potentially be much greater than the sample size n. In contrast to the existing functional data classifiers, the proposed unified model performs feature selection and classification simultaneously. The challenge arises from the complex inter-correlation structures among multiple functional processes, and at meanwhile without any assumptions on the distribution of these processes. Simulation study and real data application are carried out to demonstrate its favorable performance.

Effect Sizes and Replicability in Brain-Wide Association Studies

Simon Vandekar, Vanderbilt University Medical Center

Abstract: Several recent studies showed that thousands of study participants are required to obtain replicable results using brain-wide association studies (BWAS) because effect sizes are much smaller than those reported in smaller studies. Here, we perform analyses and meta-analyses of a robust effect size index (RESI) using 63 longitudinal and cross-sectional magnetic resonance imaging studies from the Lifespan Brain Chart Consortium (77,695 total scans) to demonstrate that optimizing study design is critical for increasing standardized effect sizes and replicability in BWAS. A meta-analysis of brain volume associations with age indicates that BWAS with larger covariate variance have larger effect size estimates and that the longitudinal studies we examined have systematically larger standardized effect sizes than cross-sectional studies. We propose a cross-sectional RESI to adjust for the systematic difference in effect sizes between cross-sectional and longitudinal studies that allows investigators to quantify the benefit of conducting their study longitudinally. Analyzing age effects on global and regional brain measures from the United Kingdom Biobank and the Alzheimer's Disease Neuroimaging Initiative, we show that modifying longitudinal study design through sampling schemes can increase between-subject variability and adding a single additional longitudinal measurement per subject improves effect sizes. However, evaluating these longitudinal sampling schemes on cognitive, psychopathology, and demographic associations with structural and functional brain outcome measures in the Adolescent Brain and Cognitive Development dataset shows that commonly used longitudinal models can, counterintuitively, reduce effect sizes. We demonstrate that the benefit of conducting longitudinal studies depends on the strengths of the between- and within-subject associations of the brain and non-brain measures. Explicitly modeling between- and within-subject effects avoids conflating the effects and allows optimizing effect sizes for them separately. These findings underscore the importance of considering study design features to improve the replicability of BWAS.

Bayesian Product Mixture Models for Multi-Scale Clustering of High-Dimensional Imaging Data

Suprateek Kundu, University of Texas at MD Anderson Cancer Center

Abstract: Bayesian non-parametric methods based on Dirichlet process mixtures have seen tremendous success in various domains and are appealing in being able to borrow information by clustering samples that share identical parameters. However, such methods can face hurdles in heterogeneous settings where objects are expected to cluster only along a subset

of axes or where clusters of samples share only a subset of identical parameters. We overcome such limitations by developing a novel class of product of Dirichlet process location-scale mixtures that enable independent clustering at multiple scales, which result in varying levels of information sharing across samples. First, we develop the approach for independent multivariate data. Subsequently we generalize it to multivariate time-series data under the framework of multi-subject Vector Autoregressive (VAR) models that is our primary focus, which go beyond parametric single-subject VAR models. We establish posterior consistency and develop efficient posterior computation for implementation. Our research generalizes the literature on Bayesian local partition models for lower dimensional independent functional data to the case of multivariate time-series data. Extensive numerical studies involving VAR models show distinct advantages over competing methods, in terms of estimation, clustering, and feature selection accuracy. Our resting state fMRI analysis from the Human Connectome Project reveals biologically interpretable connectivity differences between distinct intelligence groups, while another air pollution application illustrates the superior forecasting accuracy compared to alternate methods.

A Bayesian Model for Erroneous Links in Functional Brain Networks

Panpan Zhang, Vanderbilt University Medical Center

Abstract: Erroneous links are common in network analysis. In a complex network, links can be unobserved (false negative) due to detection limitations and technical challenges during data collection. On the other hand, spurious links (false positive) may emerge due to data noise. Inference based on networks with the presence of erroneous links may be biased in network-based downstream analysis. We propose a Bayesian model that considers erroneous links as misclassified binary outcomes, followed by an MCMC-based algorithm for Bayesian inference. We empirically justify the model via an extensive simulation study and apply the proposed method to brain network data in Alzheimer's disease.

Use of Real-World Data Enhancing Clinical Development (IS-8)

Chair: Ming-Hui Chen

Proposer: Haitao Chu, Pfizer Inc.

Room: McHugh 101

Presenters: Claire Ruixuan Zhu; Guohui Wu; Birol Emir

Discounting Individuals Power Prior: A Bayesian Approach for Borrowing Information From External Data

Claire Zhu, University of North Carolina Chapel Hill

Abstract: An increasingly popular approach in clinical trials involves leveraging external data to augment the data of randomized controlled trials. While designing a study that uses concurrent data allows for direct treatment comparisons, recruiting a sufficient number of patients, particularly in rare disease areas, can be difficult and time consuming. External borrowing of data offers a cost-effective alternative to enrich trial data, but potential non-exchangeability with current data poses challenges. We propose the discounted individual power prior (DIPP). The DIPP is an extension of the power prior that incorporates a local individualized discounting parameter along with the global discounting parameter. The local discounting assesses exchangeability at the individual level, enabling dynamic borrowing. We elicit a spike-and-slab type prior with the probability equal to the assumed degree of exchangeability between external and concurrent trial data, which is itself treated as random. We illustrate the performance of the DIPP by comparing it to the power prior in simulation exercises. We also apply our method in a rigorous analysis of real-world clinical trial data.

Practical Application of Bayesian RWD Borrowing for Clinical Trials

Guohui Wu, Amgen

Abstract: Bayesian dynamic borrowing provides a flexible way to integrate external RWD to the design and analysis of a current trial to improve efficiency. External RWD resources include historical studies, trials from similar population, and data from a different medical product in the same class. Borrowing can be from the control arm only, or from both the control arm and treatment arm then combine the two posteriors to obtain the treatment effectiveness. Borrow-

ing can also be done directly on the treatment effect scale, such as the hazard ratio, odds ratio, and mean difference and the methods include the power prior, commensurate prior, mixture prior, meta-analytic predictive prior, Bayesian hierarchical model, and propensity score integrated priors. An appropriate statistical approach needs to leverage borrowing of information while considering the heterogeneity between the external and current data. Despite various methods to use, we will focus on power prior and Bayesian hierarchical model, for which we develop well-customized variational Bayes algorithm to achieve efficient inference. Examples will be provided to demonstrate the efficiency of our methods.

Case Example for Augmenting A Randomized Control Trial Control Arm with Real-World Data

Birol Emir, Pfizer inc

Abstract: Randomized controlled trials (RCT) are generally accepted as one of the best ways to remove bias in clinical studies. However, an RCT may not always be practical or ethical. In the evolving landscape of healthcare research, the integration of Real-World Data (RWD) with traditional clinical study designs has emerged as a critical frontier. Hybrid clinical trials offer one way to augment an RCT, but with potential biases and confounding. Methodological developments to better control of bias and confounding as well as improve casual inferences will be discussed. We cover some of the prominent methods and outline plans for the application in a case example with an augmented control arm using an external set of patients' data (RWD) to boost sample size.

Novel Statistical Modeling of Correlated Data (IS-29)

Chair: Fernanda Schumacher

Proposer: Fernanda Schumacher and Victor Hugo Lachos, The Ohio State University (and UCONN)

Room: McHugh 201

Presenters: Jalmar Carrasco; Carina Brunehilde Pinto Da Silva; Kelin Zhong; Brisilda Ndreka

Hierarchical and Multivariate Regression Models to Fit Correlated Asymmetric Positive Continuous Outcomes

Jalmar M F Carrasco, Federal University of Bahia, Brazil

Abstract: In the extant literature, hierarchical models typically assume a flexible distribution for the random-effects. The random-effects approach has been used in the inferential procedure of the generalized linear mixed models . In this paper, we propose a random intercept gamma mixed model to fit correlated asymmetric positive continuous outcomes. The generalized log-gamma (GLG) distribution is assumed as an alternative to the normality assumption for the random intercept. Numerical results demonstrate the impact on the maximum likelihood (ML) estimator when the random-effect distribution is misspecified. The extended inverted Dirichlet (EID) distribution is derived from the random intercept gamma-GLG model that leads to the EID regression model by supposing a particular parameter setting of the hierarchical model. Monte Carlo simulation studies are performed to evaluate the asymptotic behavior of the ML estimators from the proposed models. Analysis of diagnostic methods based on quantile residual and COVARATIO statistic are used to assess departures from the EID regression model and identify atypical subjects. Two applications with real data are presented to illustrate the proposed methodology.

Antedependence Models for Longitudinal Data with Skewed and Heavy-Tailed Distributions

Carina Brunehilde Pinto Da Silva, UVA/UNICAMP/OSU

Abstract: Handling longitudinal data makes data analysis difficult since classical analysis procedures commonly fail to capture the autocorrelation between observations from the same subject. Moreover, an extra challenge for researchers is added when observations are nonstationary. In this work, we develop a parsimonious parametric class of models to analyze nonstationary longitudinal data with no restrictions in variances or same-lag correlations behavior along the time points, the antedependence models. In most investigations considering these models, the data is generally assumed to be normally distributed, although this assumption is not reasonable in cases of skewness, heavy tails, or multimodality. So, in this study, we propose to overcome these intricacies by extending antedependence modeling to nonstationary longitudinal data, assuming more general classes of distributions to the innovations, such as skew-normal and skew-t

ones. An advantage of using these distributions is that they are in the Scale Mixture Skew Normal family and can be written in a hierarchical form, which is convenient for their computational implementation and estimation procedures. Concerning the estimation, we adopted a Bayesian approach via Hamiltonian Monte Carlo using the STAN software due to its useful convergence efficiency and user-friendly syntax. Simulation studies were conducted to analyze important aspects, such as Markov chain convergences, parameter recovery, and the choice of priors presenting promising results.

Bayesian Analysis of Censored Linear Mixed-Effects Models for Heavy-Tailed Irregularly Observed Repeated Measures

Kelin Zhong, University of Connecticut

Abstract: HIV RNA viral load measures are often subjected to some upper or lower detection limit, depending on the quantification assays. Hence, the responses are either left- or right-censored. Censored mixed-effects models are routinely used to analyze this type of data and are based on the conditionally independent normal assumption for random errors. However, these assumptions might not provide robust inference in the presence of atypical observations. In this work, we develop a Bayesian analysis of censored linear models, replacing the Gaussian assumptions with the flexible class of scale mixture of normal distributions. A damped exponential correlation structure is considered in the random error to address the autocorrelation existing among the within-subject irregularly observed measures. Stan's default No-U-Turn sampler is utilized to obtain posterior simulations. The proposed methods are illustrated with intensive simulations and the analysis of two real AIDS case studies.

Homophily through Skewed Link Functions in Bayesian Network Models: Estimating Peer Influence

Brisilda Ndreka, University on Connecticut

Abstract: This study investigates the effectiveness of the asymmetric link function in modeling relationship probabilities among actors in network models. To distinguish between influence effects and homophily in a social analysis, we specifically designed a dynamic Bayesian framework. Particularly, the study adopts the skew link function to explain the complex structure of evolving networks. Through simulation studies, the substantial impact of tie proportions on accurately estimating peer influence has been demonstrated by comparing models using symmetric and skewed links. In addition, we present an illustrative analysis using longitudinal data on students' friendship networks and their engagement in social and behavioral health.

Recent Advances in High-Dimensional Structured Data Modeling (IS-39)

Chair: Yuping Zhang

Proposer: Yuping Zhang, University of Connecticut

Room: McHugh 301

Presenters: Yang Ning; Zhengqing Ouyang; Julio Castrillon; Yuping Zhang

Analysis of RNA Higher-Order Structures and Interactions

Zhengqing Ouyang, University of Massachusetts Amherst

Abstract: RNA fate and function are informed by their structures. The higher-order structures of RNAs remain largely unknown. By analyzing experimental data from high-throughput sequencing, KARR-seq, we characterize RNA–RNA interactions and detect higher-order RNA structures in living cells. Benchmarking with known structures demonstrates high sensitivity and accuracy of our approach. We further show that translation process represses mRNA higher-order structure globally.

Analysis of RNA Higher-Order Structures and Interactions

Zhengqing Ouyang, University of Massachusetts Amherst

Abstract: RNA fate and function are informed by their structures. The higher-order structures of RNAs remain largely unknown. By analyzing experimental data from high-throughput sequencing, KARR-seq, we characterize RNA–RNA interactions and detect higher-order RNA structures in living cells. Benchmarking with known structures demonstrates high

sensitivity and accuracy of our approach. We further show that translation process represses mRNA higher-order structure globally.

Uncertainty Quantification for Non-Linear Stochastic Networks

Julio Enrique Castrillon Candas, Boston University

Abstract: Despite the popularity of stochastic networks, they are not well understood in the context of non-linear constraints between the variables of the graph. Due to the non-linearity, it is difficult to quantify the statistics of the graph analytically, thus leaving computational methods as the only viable option. In this talk, concepts from uncertainty quantification for partial differential equations and numerical analysis are introduced in the context of efficient evaluation of high-dimensional stochastic Newton iterates with applications to non-linear networks. In particular, complex analytic regularity theory is developed for the solution with respect to the random variables. This justifies the application of stochastic collocation with sparse grids for the computation of statistical quantities. Convergence rates are derived and shown to be sub-exponential or algebraic with respect to the number of realizations of random perturbations. Furthermore, due to the accuracy of the method, sparse grids are also well-suited for computing low probability events with high confidence. This approach is tested on the non-linear power flow equations for electric networks. Numerical experiments on the 39-bus New England power system model with large stochastic loads are consistent with the theoretical convergence rates. Moreover, compared to the Monte Carlo method, our approach is at least 100,000,000,000 times faster for the same accuracy.

Translocation Detection From Hi-C Data via Scan Statistics

Yuping Zhang, University of Connecticut

Abstract: Recent Hi-C technology enables more comprehensive chromosomal conformation research, including the detection of structural variations, especially translocations. In this paper, we formulate the interchromosomal translocation detection as a problem of scan clustering in a spatial point process. We then develop TranScan, a new translocation detection method through scan statistics with the control of false discovery. The simulation shows that TranScan is more powerful than an existing sophisticated scan clustering method, especially under strong signal situations. Evaluation of TranScan against current translocation detection methods on realistic breakpoint simulations generated from real data suggests better discriminative power under the receiver-operating characteristic curve. Power analysis also highlights TranScan's consistent outperformance when sequencing depth and heterozygosity rate is varied. Comparatively, Type I error rate is lowest when evaluated using a karyotypically normal cell line. Both the simulation and real data analysis indicate that TranScan has great potentials in interchromosomal translocation detection using Hi-C data.

Exploring New Frontiers in Quantile Regression (IS-51)

Chair: Kun Chen

Proposer: Kun Chen, University of Connecticut

Room: McHugh 305

Presenters: Jun Jin; Ting Zhang; Haim Bar; Yuwen Gu

Transfer Learning with Large-Scale Quantile Regression

Jun Jin, University of Connecticut

Abstract: Quantile regression is increasingly encountered in modern big data applications due to its robustness and flexibility. We consider the scenario of learning the conditional quantiles of a specific target population when the available data may go beyond the target and be supplemented from other sources that possibly share similarities with the target. A crucial question is how to properly distinguish and use useful information from other sources to improve the quantile estimation and inference at the target. We develop transfer learning methods for high-dimensional quantile regression by detecting informative sources whose models are similar to the target and using them to improve the target model. We show that under reasonable conditions, the detection of the informative sources based on sample splitting is consistent. Compared to the naive estimator with only the target data, the transfer learning estimator achieves a much lower error rate as a function of the sample sizes, the signal-to-noise ratios, and the similarity measures

among the target and the source models. Extensive simulation studies demonstrate the superiority of our proposed approach. We apply our methods to tackle the problem of detecting hard-landing risk for flight safety and show the benefits and insights gained from transfer learning of three different types of airplanes: Boeing 737, Airbus A320, and Airbus A380.

High-Quantile Regression for Tail-Dependent Time Series

Ting Zhang, University of Georgia

Abstract: Quantile regression is a popular and powerful method for studying the effect of regressors on quantiles of a response distribution. However, existing results on quantile regression were mainly developed for cases in which the quantile level is fixed, and the data are often assumed to be independent. Motivated by recent applications, we consider the situation where (i) the quantile level is not fixed and can grow with the sample size to capture the tail phenomena, and (ii) the data are no longer independent, but collected as a time series that can exhibit serial dependence in both tail and non-tail regions. To study the asymptotic theory for high-quantile regression estimators in the time series setting, we introduce a tail adversarial stability condition, which had not previously been described, and show that it leads to an interpretable and convenient framework for obtaining limit theorems for time series that exhibit serial dependence in the tail region, but are not necessarily strongly mixing. Numerical experiments are conducted to illustrate the effect of tail dependence on high-quantile regression estimators, for which simply ignoring the tail dependence may yield misleading p-values.

Quantile Regression Modelling via Location and Scale Mixtures of Normal Distributions

Haim Bar, University of Connecticut

Abstract: We show that the estimating equations for quantile regression can be solved using a simple EM algorithm in which the M-step is computed via weighted least squares, with weights computed at the E-step as the expectation of independent generalized inverse-Gaussian variables. We compute the variance-covariance matrix for the quantile regression coefficients using a kernel density estimator that results in more stable standard errors than those produced by existing software. A natural modification of the EM algorithm that involves fitting a linear mixed model at the M-step extends the methodology to mixed effects quantile regression models. In this case, the fitting method can be justified as a generalized alternating minimization algorithm. Obtaining quantile regression estimates via the weighted least squares method enables model diagnostic techniques similar to the ones used in the linear regression setting.

Fastkqr: A Fast Algorithm for Kernel Quantile Regression

Yuwen Gu, University of Connecticut

Abstract: Quantile regression is a powerful tool for robust and heterogeneous learning that has seen applications in a diverse range of applied areas. Its broader application, however, is often hindered by the substantial computational demands arising from the nonsmooth quantile loss function. We introduce a novel algorithm named fastkqr, which significantly advances the computation of quantile regression in reproducing kernel Hilbert spaces. The essence of fastkqr is a finite smoothing algorithm that magically produces exact regression quantiles, rather than approximations. To further accelerate the algorithm, we equip fastkqr with an innovative spectral technique that carefully reuses matrix computations. In addition, we extend fastkqr to solve a flexible kernel quantile regression with a data-driven crossing penalty. The new method addresses the interpretability issue with quantile regression where fitted quantile curves at multiple levels are often crossing in a finite sample. Extensive simulations and real applications show that fastkqr achieves the same accuracy as the state-of-the-art algorithms but can be an order of magnitude faster.

New Methods for Robust Inference and Selection (IS-58)

Chair: Omar Melikechi

Proposer: Omar Melikechi, Harvard University

Room: McHugh 202

Presenters: Tim Barry; Maryclare Griffin; Omar Melikechi; Neil Spencer

Robust Inference for Single-Cell CRISPR

Screens via Resampling Score Statistics

Timothy Barry, Harvard University

Abstract: Single-cell CRISPR screens (perturb-seq) link genetic perturbations to phenotypic changes in individual cells. The most fundamental task in perturb-seq analysis is to test for association between a perturbation and a count outcome, such as gene expression. We conduct the first-ever comprehensive benchmarking study of association testing methods for low multiplicity-of-infection (MOI) perturb-seq data, finding that existing methods produce excess false positives. We conduct an extensive empirical investigation of the data, identifying three core analysis challenges: sparsity, confounding, and model misspecification. Finally, we develop an association testing method that resolves these analysis challenges and demonstrates improved calibration and power. The proposed method is based on the novel and statistically principled technique of resampling negative binomial score statistics.

A Simple Approach for Local and Global Variable Importance in Nonlinear Regression Models

Maryclare Griffin, University of Massachusetts Amherst

Abstract: The ability to interpret machine learning models has become increasingly important as their usage in data science continues to rise. Most current interpretability methods are optimized to work on either (*i*) a global scale, where the goal is to rank features based on their contributions to overall variation in an observed population, or (*ii*) the local level, which aims to detail on how important a feature is to a particular individual in the data set. In this work, a new operator is proposed called the "GlObal And Local Score" (GOALS): a simple *post hoc* approach to simultaneously assess local and global feature variable importance in nonlinear models. Motivated by problems in biomedicine, the approach is demonstrated using Gaussian process regression where the task of understanding how genetic markers are associated with disease progression both within individuals and across populations is of high interest. Detailed simulations and real data analyses illustrate the flexible and efficient utility of GOALS over state-of-the-art variable importance strategies.

Integrated Path Stability Selection

Omar Melikechi, Harvard University

Abstract: Stability selection is a widely used method for improving the performance of feature selection algorithms. However, it can be highly conservative, resulting in low sensitivity. Further, existing bounds on the expected number of false positives, E(FP), are relatively loose, making it difficult to know how many false positives to expect in practice. In this talk, I will introduce a novel method for stability selection based on integrating the stability paths rather than maximizing over them. This yields a tighter bound on E(FP), resulting in a feature selection criterion that has higher sensitivity in practice and is better calibrated in terms of matching the target E(FP). Our proposed method requires the same amount of computation as the original stability selection algorithm, and only requires the user to specify one input parameter, a target value for E(FP). I will discuss theoretical bounds on performance, and demonstrate the method on simulations and real data from cancer gene expression studies.

Robust Bayesian Model Selection for Network Data

Neil A. Spencer, University of Connecticut

Abstract: The asymptotic behavior of Bayesian model selection is well-understood in the context of independent and identically distributed data. As the number of observations increases, the posterior concentrates on whichever candidate model is closest to the truth, with closeness determined by the Kullback-Leibler divergence. In this work, I extend these results to node exchangeable network data. My contributions include: (1) establishing criteria for posterior concentration, (2) determining an appropriate notion of "closeness" to the truth, (3) identifying scenarios where Bayesian model selection is unstable, and (4) proposing an extension of BayesBag to network data to address this instability.

Novel Statistical and Machine Learning Applications with Complicated Data (IS-64)

Chair: TBD

Proposer: Kun Chen, University of Connecticut

Room: McHugh 205

Presenters: Haiyan Su; Hon Kiu To; Shiya Cao; Timothy Becker; Ying Li

Survival Prediction in Amyotrophic Lateral Sclerosis Using Deep Learning

Haiyan Su, Montclair State University

Abstract: Amyotrophic lateral sclerosis is a progressive neurodegenerative disease that affects nerve cells in the brain and spinal cord, affecting approximately 31,000 people in the United States. The FDA has approved several drugs for ALS that may reduce the rate of decline, or help manage symptoms. However, there is currently no known treatment that stops or reverses the progression of ALS. A more accurate survival prediction may help for future clinical trial design to understand the progression of ALS and further prolong survival. In this study, we investigated the possibility of using deep learning to better predict survival for ALS patients using data from the PRO-ACT database. After developing the deep neural network model, it was compared with two models in the literature: the traditional statistical model using Cox Proportional Hazard (CPH) with ElasticNet, and a reliable machine learning model, Gradient Boosting Machine (GBM). The comparison showed that deep learning model is comparable to GBM, and both models are superior to the CPH, in terms of prediction accuracy.

Accuracy and Fairness in The Use of Facial Recognition Technology by Law Enforcement

Hon Kiu To, University of Pennsylvania

Abstract: With the increasing prevalence of facial recognition technology (FRT), law enforcement officers and legal practitioners need to understand its accuracy when applied to real-world images, as opposed to the high-quality images typically used for testing. Without this knowledge, they might trust the algorithm more or less than they should. This could lead to misclassifications, and thus miscarriages of justice. Our study aims to evaluate the accuracy and fairness of a specific FRT, offering insights for various stakeholders such as police departments, defense attorneys, judges, researchers, and society at large. Using StyleGAN3, we generate high-quality synthetic faces labeled with FairFace for demographic attributes. Subsequently, we simulate real-world conditions by manipulating illumination and resolution to create low-quality images. These images are then subjected to facial recognition tasks using Deepface, which incorporates state-of-the-art models employing the ArcFace loss function. Our results show that the image qualities have a significant impact on the accuracy and fairness of FRT. This emphasizes the need for further research to ensure the appropriate implementation and interpretation of FRT results within the criminal justice system.

Model Interpretation after Using Random Projections: An Applied Study on Travel Disability Data

Shiya Cao, Smith College

Abstract: The National Household Travel Survey (NHTS) asks respondents whether they have a medical condition "that makes it difficult to travel outside of home", which is defined as travel disability in this research. The NHTS allows us to investigate the effects of disability on travel behavior, however, it may release some sensitive medical conditions and travel data. We use a differential privacy algorithm – random projection to get a random dataset that contains the summary statistics of the sample dataset so useful aggregate information can be released and used for the intended purposes, while the privacy of the individuals in the sample dataset is preserved. The main idea of this differential privacy algorithm is to use random projection to project a sample dataset (n by p) to a random dataset (k by p). We fit a linear regression model for the random dataset and compare the statistics of interest of the random dataset with those of the sample dataset. With this differential privacy algorithm, we can examine the accuracy of our random projection compared to the original sample and then make statements about statistics of interest of the true population. This differential privacy algorithm can be applied to other disability datasets that contain sensitive information.

A Transparent and Explainable Ensemble Method for Classification of Human Sequence Variation

Timothy, Connecticut College

Abstract: Ensemble methods achieve significant increases in accuracy using multiple specialized models as input, but often lack the transparency around the decision process that is needed for interpreting human genetic results. FusorSV is a black box ensemble method that takes sequence variation data that is processed by multiple weak models as its input. The questions of which weak models to use, and how to

combine them are answered by a specially designed metric and data fusion process that estimates a likelihood to maximize accuracy. When the likelihood exceeds a threshold found during fitting, it will suggest a decision which is reliable and up to 80% accurate according to follow up in vitro validation. We provide some needed updates to this method by calibrating the likelihood estimates to a secondary dataset and make modifications to keep track of data rows that were used in the maximization of the ensemble. When the model is then used for inference, we can generate the path in the supporting evidence for each decision which yields a more interpretable result.

Uncertainty Quantification in Machine Learning for Glass Transition Temperature Prediction of Polymers

Ying Li, University of Wisconsin–Madison

Abstract: Machine learning (ML) has become an important technique in materials science, markedly accelerating the discovery and design of novel materials, and concurrently lowering the burden of experimental costs. Uncertainty quantification (UQ) plays a pivotal role in the accurate prediction and innovative design of novel materials through ML techniques. In this study, we perform a comprehensive evaluation of six UQ methods in ML, including ensemble, Gaussian process regression (GPR), Monte Carlo dropout (MCD), Mean-variance estimation (MVE), Bayesian neural network (BNN) and Evidential deep learning (EDL), for predictions on the glass transition temperature of polymers. We assess the accuracy and performance of these UQ methods using three metrics, including Spearman's rank correlation coefficient, calibration and sparsification, offering a substantial reference for data-driven polymer design. Our analysis encompasses test data, out-of-distribution data from experiments and molecular dynamics simulations, and high-polymer data for UQ analysis of ML predictions. The results indicate that ML models are robust and effective in predicting polymer values for testing and experimental data. However, correlating actual errors with uncertainties (standard deviations) poses a significant challenge, with ML models frequently exhibiting overconfidence with low uncertainties. Moreover, the accuracy of ML predictions improves when the data with large uncertainties are excluded, suggesting a potential strategy for refining ML model's performance.

Student Paper 1 (S1)

Chair: Qi Zhang

Organizer: Neil Spencer

Room: McHugh 306

A Bayesian Model of Underreporting for Sexual Assault on College Campuses

Casey Bradshaw, Columbia University

Co-authors: David Blei

Abstract: In an effort to quantify and combat sexual assault, US colleges and universities are required to disclose the number of reported sexual assaults on their campuses each year. However, many instances of sexual assault are never reported to authorities, and consequently the number of reported assaults does not fully reflect the true total number of assaults that occurred; the reported values could arise from many combinations of reporting rate and true incidence. In this paper we estimate these underlying quantities via a hierarchical Bayesian model of the reported number of assaults. We use informative priors, based on national crime statistics, to act as a tiebreaker to help distinguish between reporting rates and incidence. We outline a Hamiltonian Monte Carlo (HMC) sampling scheme for posterior inference regarding reporting rates and assault incidence at each school, and apply this method to campus sexual assault data from 2014-2019. Results suggest an increasing trend in reporting rates for the overall college population during this time, while individual school-level results are markedly diverse.

Adapt: Analysis of Microbiome Differential Abundance by Pooling Tobit Models

Mukai Wang, University of Michigan

Co-authors: Simon Fontaine, Hui Jiang, Gen Li

Abstract: Microbiome differential abundance analysis remains a challenging problem despite multiple methods proposed in the literature. The excessive zeros and compositionality of metagenomics data are two main challenges for differential abundance analysis. We propose a novel method called "analysis of differential abundance by pooling Tobit models" (ADAPT) to overcome these two challenges. ADAPT uniquely treats zero counts as left-censored observations to facilitate computation and enhance interpretation. ADAPT also encompasses a theoretically justified way of selecting non-differentially abundant

microbiome taxa as a reference for hypothesis testing. We generate synthetic data using independent simulation frameworks to show that ADAPT has more consistent false discovery rate control and higher statistical power than competitors. We use ADAPT to analyze 16S rRNA sequencing of saliva samples and shotgun metagenomics sequencing of plaque samples collected from infants in the COHRA2 study. The results provide novel insights into the association between the oral microbiome and early childhood dental caries.

Double Trouble: Predicting New Variant Counts Across Two Heterogeneous Populations

Yunyi Shen, MIT

Co-authors: Lorenzo Masoero, Joshua G. Schraiber, Tamara Broderick

Abstract: Collecting genomics data across multiple heterogeneous populations (e.g., across different cancer types) has the potential to improve our understanding of disease. Despite sequencing advances, though, resources often remain a constraint when gathering data. So it would be useful for experimental design if experimenters with access to a pilot study could predict the number of new variants they might expect to find in a follow-up study: both the number of new variants shared between the populations and the total across the populations. While many authors have developed prediction methods for the single-population case, we show that these predictions can fare poorly across multiple populations that are heterogeneous. We prove that, surprisingly, a natural extension of a state-of-the-art single-population predictor to multiple populations fails for fundamental reasons. We provide the first predictor for the number of new shared variants and new total variants that can handle heterogeneity in multiple populations. We show that our proposed method works well empirically using real cancer and population genetics data.

Maximum A Posteriori Inference for Factor Graphs via Benders' Decomposition

Harsh Vardhan Dubey, University of Massachusetts Amherst

Co-authors: Ji Ah Lee, Patrick Flaherty

Abstract: Many Bayesian statistical inference problems come down to computing a maximum a-posteriori (MAP) assignment of latent variables. Yet, standard methods for estimating the MAP assignment do not have a finite time guarantee that the algorithm has converged to a fixed point. Previous research has found that MAP inference can be represented in dual form as a linear programming problem with a non-polynomial number of constraints. A Lagrangian relaxation of the dual yields a statistical inference algorithm as a linear programming problem. However, the decision as to which constraints to remove in the relaxation is often heuristic. We present a method for maximum a-posteriori inference in general Bayesian factor models that sequentially adds constraints to the fully relaxed dual problem using Benders' decomposition. Our method enables the incorporation of expressive integer and logical constraints in clustering problems such as must-link, cannot-link, and a minimum number of whole samples allocated to each cluster. Using this approach, we derive MAP estimation algorithms for the Bayesian Gaussian mixture model and latent Dirichlet allocation. Empirical results show that our method produces a higher optimal posterior value compared to Gibbs sampling and variational Bayes methods for standard data sets and provides certificate of convergence.

A Bayesian Joint Hierarchical Modeling Approach to Determining The Impact of Flavoring on The Addictiveness of Cigars

Zoe Gibbs McBride, University of Connecticut

Co-authors: Xiaojing Wang, Timothy Moore, Erin Mead-Morse

Abstract: Recent debates surrounding the use of flavoring in tobacco products have led to the use of hypothetical purchase tasks (HPT) in an experimental setting to quantify the impact of flavoring on tobacco product addictiveness. We introduce a novel method of simultaneously estimating the intensity and elasticity of demand, as well as their response to treatment, from HPTs using Bayesian joint hierarchical mixed modeling. While previous analyses estimated treatment effects assuming the intensity and elasticity of demand are known quantities, our method allows us to account for the uncertainty associated with the demand quantities and the variation in individual responses. We apply this model to a cross-over design where smokers are assigned to smoke flavored and unflavored cigars. We find that while the intensity of demand is not significantly different between flavored and unflavored cigars, elasticity tends to be lower for flavored cigars than unflavored cigars. This supports previous assertions that flavored cigars may be more addictive than unflavored cigars.

Parallel Session 2 | 02:00 PM - 03:40 PM, May 22

Frontiers in Financial Mathematics (IS-1)

Chair: Oleksii Mostovyi

Proposer: Oleksii Mostovyi, University of Connecticut

Room: McHugh 206

Presenters: Oleksii Mostovyi; Bahman Angoshtari; Scott Robertson; Maxim Bichuch

An Approach to The Greeks for Indifference Pricing

Oleksii Mostovyi, University of Connecticut

Abstract: We consider the problem of sensitivity of the indifference pricing to the dynamics of the underlying assets. In the context of arbitrage-free pricing (AFP), such sensitivities are known as the Greeks. Here, in multidimensional semimartingale settings of incomplete models, we develop the computations of the Greeks and the associated trading strategies for indifference pricing in the sense of Davis. Unlike the traditional AFP, e.g., in the Black-Scholes model, where the Greeks represent the sensitivity of a linear pricing problem to perturbations of the stock price dynamics, as indifference prices are given via solutions to (non-linear) optimization problems, their sensitivities to perturbations of model parameters, that is the Greeks, are also represented by value functions of (auxiliary quadratic) optimization problems, which we introduce too. The proposed approach also allows for the hedging of nonreplicable contingent claims, in contrast to the Greeks for AFP-based hedging in incomplete markets, where the AFPs form intervals, and their derivatives are not defined in the usual sense. We illustrate the results with positive examples.

Rank-Dependent Predictable Forward Performance Processes

Bahman Angoshtari, University of Miami

Abstract: Predictable forward performance processes (PFPPs) are stochastic optimal control frameworks for an agent who controls a randomly evolving system but can only prescribe the system dynamics for a short period ahead. This is a common scenario in which a controlling agent frequently re-calibrates her model. We introduce a new class of PFPPs based on rank-dependent utility, generalizing existing models that are based on expected utility theory (EUT). We establish existence of rank-dependent PFPPs under a conditionally complete market and exogenous probability distortion functions which are updated periodically. We show that their construction reduces to solving an integral equation that generalizes the integral equation obtained under EUT in previous studies. We then propose a new approach for solving the integral equation via theory of Volterra equations. We illustrate our result in the special case of conditionally complete Black-Scholes model.

Equilibrium with Heterogeneous Information Flows

Scott Robertson, Questrom School of Business, Boston University

Abstract: We study a continuous time economy where throughout time, insiders receive private signals regarding the risky assets' terminal payoff. We prove existence of a partial communication equilibrium where, at each private signal time, the public receives a signal of the same form as the associated insider, but of lower quality. This causes a jump in both the public information flow and equilibrium asset price. The resultant markets, while complete between each jump time, are incomplete over each jump. After establishing equilibrium for a finite number of private signal times, we consider the limit as the private signals become more and more frequent. Under appropriate scaling we prove convergence of the public filtration to the natural filtration generated by both the fundamental factor process X and a continuous time process J taking the form $J(t) = X(1) + Y(t)$ where $X(1)$ is the asset terminal payoff and Y an independent Gaussian process.

Advanced Spatial Learning Methods for Biomedical Applications (IS-6)

Chair: Garrett Frady

Proposer: Shariq Mohammed, Boston University

Room: McHugh 301

Presenters: Aritra Halder; Ying Ma; Jianing Wang;

Lukas Weber

Bayesian Modeling with Spatial Curvature Processes

Aritra Halder, Drexel Univeristy

Abstract: Spatial process models are widely used for modeling point-referenced variables arising from diverse scientific domains. Analyzing the resulting random surface provides deeper insights into the nature of latent dependence within the studied response. We develop Bayesian modeling and inference for rapid changes on the response surface to assess directional curvature along a given trajectory. Such trajectories or curves of rapid change, often referred to as wombling boundaries, occur in geographic space in the form of rivers in a flood plain, roads, mountains or plateaus or other topographic features leading to high gradients on the response surface. We demonstrate fully model based Bayesian inference on directional curvature processes to analyze differential behavior in responses along wombling boundaries. We illustrate our methodology with a number of simulated experiments followed by multiple applications featuring the Boston Housing data; Meuse river data; and temperature data from the Northeastern United States. Supplementary materials for this article are available online.

Accurate and Efficient Integrative Reference-Informed Spatial Domain Detection for Spatial Transcriptomics

Ying Ma, Brown University

Abstract: Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and increasingly large, offering unprecedented opportunities to characterize the spatial and functional organization of complex tissues. Here, we introduce a computational method, IRIS, that characterizes the spatial organization of complex tissues through accurate and efficient detection of spatial domains. IRIS uniquely leverage the widespread availability of single-cell RNA-seq data for reference-informed spatial domain detection, integrates multiple SRT tissue slices jointly while explicitly considering correlation both within and across slices, produces biologically interpretable spatial domains, and benefits from multiple algorithmic innovations for highly scalable computation. We demonstrate the advantages of IRIS through in-depth analysis of six SRT datasets from different technologies across various tissues, species, and spatial resolutions. In these applications, IRIS attains an unprecedented 58% 1,083% accuracy gain over existing methods in the gold standard dataset with known ground truth. Furthermore, IRIS is 8.5 134.7 times faster than existing methods in moderate-sized datasets and is the only method applicable to large-scale SRT datasets, including the very recent stereo-seq and 10x Xenium. As a result, IRIS uncovers the fine-scale structures of brain regions, reveals the spatial heterogeneity of distinct tumor microenvironments, and characterizes the structural changes of the seminiferous tubules in the testis associated with diabetes, all at a speed and accuracy unachievable by existing approaches.

A Two-Stage Spatial Capture-Recapture Approach for Prevalence Estimation Using Administrative Health Data

Jianing Wang, Massachusetts General Hospital, Harvard Medical School

Abstract: Accurately estimating the sizes of hidden populations across geographic groups is crucial for policymaking. Capture-recapture methods (CRC), initially used in ecology, have been adapted in epidemiology as an indirect method to improve disease prevalence estimation. However, conventional CRC methods face challenges in estimating prevalence in small areas and do not account for underlying correlations. To address these issues, we propose a two-stage Bayesian hierarchical CRC model by considering individual heterogeneity at the first stage and disease prevalence variation across space at the second stage. Simulation studies assess the models' performance in various scenarios, demonstrating improved estimate accuracy with reduced variance. Importantly, it handles data sparsity challenges and provides smoothed estimates across groups. We also apply it to the Massachusetts Public Health Data Warehouse to estimate opioid use disorder prevalence among cities. Our framework serves as a powerful tool for estimating disease burden within subgroups, and its application is essential for informing targeted interventions and developing effective public health policies.

Identification of Spatial Domains by Smoothing for Cell Type Compositional Analyses in Spatial Transcriptomics

Lukas M. Weber, Boston University, Department of Biostatistics

Abstract: Spatial transcriptomics platforms enable the measurement of transcriptome-scale gene expression

levels at spatial resolution, and have become widely applied to study spatial variation in cell type composition within tissue samples. Depending on the technological platform, the spatial resolution of the measurements may either be at molecular resolution or consist of pooled measurements from one or more cells per spatial location, and measurements may also be characterized by high levels of sparsity due to sampling variation. The identification of spatial domains consisting of tissue regions with relatively uniform cell type composition or mixtures and consistent gene expression signatures represents a key step during computational analysis workflows. Spatial domains may then be further investigated by applying tools for cell type compositional analyses. We have developed a new method, smoothclust, to identify spatial domains in spatial transcriptomics data in a computationally scalable manner, based on spatial smoothing of gene expression values followed by unsupervised clustering. We have evaluated the method using data from several technological platforms and compared against existing and baseline methods. The method is freely available as an R package from Bioconductor, and is integrated into Bioconductor-based analysis workflows for spatial transcriptomics data.

Modern Advances in Statistical Methodology for Meta-Analysis (IS-7)

Chair: Chongliang Luo

Proposer: Chongliang Luo, Washington University in St Louis

Room: McHugh 305

Presenters: Yu-lun Liu; Lifeng Lin; Chongliang Luo; Xiaokang Liu

Network Meta-Analysis Made Simple: A Composite Likelihood Approach

Yu-lun Liu, University of Texas Southwestern Medical Center

Abstract: Network meta-analysis, also known as mixed treatments comparison meta-analysis or multiple treatments meta-analysis, expands upon conventional pairwise meta-analysis by simultaneously synthesizing multiple interventions in a single integrated analysis. Despite the growing popularity of network meta-analysis within comparative effectiveness research, it comes with potential challenges. For example, within-study correlations among treatment comparisons are rarely reported in the published literature. Yet, these correlations are pivotal for valid statistical inferences. As demonstrated in earlier studies, ignoring these correlations can lead to inflated mean squared errors in estimates and inaccurate standard errors. In this talk, I will introduce a composite likelihood-based approach that guarantees accurate statistical inferences even without knowledge of these within-study correlations. The approach is computationally robust and efficient, with substantially reduced computational time compared to the state-of-the-science methods implemented in R packages. The proposed method has been evaluated through extensive simulations and applied to two important applications, including a network meta-analysis comparing interventions for primary open-angle glaucoma and another comparing treatments for chronic prostatitis and chronic pelvic pain syndrome.

Nonparametric Bayesian Approach to Treatment Ranking in Network Meta-Analysis

Lifeng Lin, University of Arizona

Abstract: Network meta-analysis is a powerful tool to synthesize evidence from independent studies and compare multiple treatments simultaneously. A critical task of performing a network meta-analysis is to offer ranks of all available treatment options for a specific disease outcome. Frequently, the estimated treatment rankings are accompanied by a large amount of uncertainty, suffer from multiplicity issues, and rarely permit possible ties of treatments with similar performance. These issues make interpreting rankings problematic as they are often treated as absolute metrics. To address these shortcomings, we formulate a ranking strategy that adapts to scenarios with high-order uncertainty by producing more conservative results. This improves the interpretability while simultaneously accounting for multiple comparisons. To admit ties between treatment effects in cases where differences between treatment effects are negligible, we also develop a Bayesian nonparametric approach for network meta-analysis. The approach capitalizes on the induced clustering mechanism of Bayesian nonparametric methods, producing a positive probability that two treatment effects are equal. We demonstrate the utility of the procedure through numerical experiments and a network meta-analysis designed to study antidepressant treatments.

Confidence Score: A Data-Driven Measure for Inclusive Systematic Reviews Considering Unpublished Preprints

Chongliang Luo, Washington University in St Louis

Abstract: COVID-19, since its emergence in December 2019, has globally impacted research. Over 360,000 COVID-19-related manuscripts have been published on PubMed and preprint servers like medRxiv and bioRxiv, with preprints comprising about 15% of all manuscripts. Yet, the role and impact of preprints on COVID-19 research and evidence synthesis remain uncertain. We propose a novel data-driven method for assigning weights to individual preprints in systematic reviews and meta-analyses. This weight termed the "confidence score" is obtained using the survival cure model, also known as the survival mixture model, which takes into account the time elapsed between posting and publication of a preprint, as well as metadata such as the number of first 2-week citations, sample size, and study type. Using 146 preprints on COVID-19 therapeutics posted from the beginning of the pandemic through April 30, 2021, we validated the confidence scores, showing an area under the curve of 0.95 (95% CI, 0.92-0.98). Through a use case on the effectiveness of hydroxychloroquine, we demonstrated how these scores can be incorporated practically into meta-analyses to properly weigh preprints. Our proposed confidence score has the potential to improve systematic reviews of evidence related to COVID-19 and other clinical conditions by providing a data-driven approach to including unpublished manuscripts.

A Generalized Method of Moments Method for Distributed Inference of Heterogeneous and Structural Missing Data

Xiaokang Liu, University of Missouri

Abstract: In multicenter biomedical research, integrating data from multiple decentralized sites provides more robust and generalizable findings due to its larger sample size and the ability to account for the heterogeneities across different sites. However, sharing individual-level data across sites is often difficult due to patient privacy concerns and regulatory restrictions. To overcome this challenge, many distributed algorithms, that fit a global model by only communicating aggregated information across sites, have been proposed. A major challenge in applying existing distributed algorithms to real-world data is that their validity often relies on the assumption that data across sites are independently and identically distributed–an assumption that is frequently violated in practice. In biomedical applications, data distributions across clinical sites can, for example, be heterogeneous. Additionally, the set of covariates available at each site may vary due to different data collection protocols. We propose a distributed inference framework for data integration in the presence of both distribution heterogeneity and data structural heterogeneity. By modeling heterogeneous and structurally missing data using density-tilted generalized method of moments, we developed a general aggregated data-based distributed algorithm that is communication-efficient and heterogeneity-aware. We establish the asymptotic properties of our estimator and demonstrate the validity of our method in finite sample settings via simulation studies. We apply our method to identify risk factors for Alzheimer's disease using data from three Alzheimer's Disease Research Centers.

Advancing Clinical Trials through Innovative Statistical Design and Analysis (IS-37)

Chair: Yeongjin Gwon

Proposer: Yeongjin Gwon, University of Nebraska Medical Center

Room: McHugh 205

Presenters: Shirin Golchi; Hayley Belli; Lorenzo Trippa; Yeongjin Gwon

An Adaptive Enrichment Design Using Bayesian Model Averaging for Selection and Threshold-Identification of Tailoring Variables

Shirin Golchi, McGill University

Abstract: Precision medicine stands as a transformative approach in healthcare, offering tailored treatments that can significantly enhance patient outcomes and reduce healthcare costs. As understanding of complex disease improves, clinical trials are designed to detect subgroups of patients with enhanced treatment effects. Biomarker-driven designs, especially adaptive enrichment designs, which enrol a general population initially and later restrict accrual based on interim results to treatment-sensitive patients, are gaining

popularity. Current practice often assumes either pre-trial knowledge of biomarkers defining treatment-sensitive subpopulations or simple, linear relationships between continuous markers and treatment effectiveness. Motivated by a trial studying available treatments in rheumatoid arthritis, we propose a Bayesian adaptive enrichment design. Our proposed design is equipped with a flexible modelling framework where the effects of continuous biomarkers are introduced using free knot B-splines. The parameters of interest are then estimated by marginalizing over the set of all possible variable combinations using Bayesian model averaging. At interim analyses, we assess whether a biomarker-defined subgroup has enhanced or reduced treatment effects, allowing for termination due to futility or efficacy and restricting future enrollment to treatment-sensitive patients. We consider categorical and continuous biomarkers, the latter of which may have complex, nonlinear relationships to the outcome and treatment effect. Using simulations, we derive the operating characteristics of our design and compare its performance to two existing approaches.

Statistical and Design Considerations for An Adaptive Stage Sequential Multiple Assignment Randomized Trial

Hayley Belli, New York University

Abstract: In the precision medicine era, there is a need to design patient-focused, pragmatic clinical trials. In this talk, we will introduce an approach for determining individualized treatment duration in a Sequential Multiple Assignment Randomized Trial (SMART). SMARTs are an adaptive design where every participant is first randomized to a treatment or control arm, similar to a classic parallel design. However, following this initial assignment, patients move through a series of stages with the option to be re-randomized to switch treatments, depending on their response to the intervention in the stage prior. The SMART framework mimics standard clinical practice in that with time patients will have the opportunity to be assigned to more effective treatments all within the rigorous experimental framework of a randomized controlled trial. However, a limitation to the SMART design is that the duration of the stages is applied uniformly to all participants and is selected by investigators in advance. Since the primary objective of this design is to arrive at an optimal set of decision rules, the duration for which to administer a treatment is an important component of the dynamic treatment regime and should be derived experimentally, rather than selected a priori. In the present work, we introduce the concept of an adaptive stage SMART design. We propose an algorithm that uses a likelihood-based approach to determine when a patient should stay on a treatment for the full stage duration or switch interventions prior to the stage end. We first derive the algorithm, and then demonstrate its performance using data from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinical Care (EMBARC) Study, a two-stage SMART design with multiple interim data points measuring the effectiveness of sertraline in 242 patients with nonpsychotic Major Depressive Disorder (Trivedi et al. 2016. Journal of Psychiatric Research, 78: 11-23). We discuss results from simulations that explore the use of limited interim data (only two or three measurements within a single stage) to determine whether and when a patient should be re-randomized. By varying the frequency and timing of these data, simulations reveal true positive rates between 70-90% and false positive rates between 20-50%. We end by discussing some practical considerations for analyzing data, implementation, and testing intervention effectiveness within the proposed adaptive stage SMART design framework.

A Two-Stage Optimal Bayesian Adaptive Design in Phase II Clinical Trials

Yeongjin Gwon, University of Nebraska Medical Center (UNMC)

Abstract: The personalized/precision medicine has accelerated the pace of scientific discovery, providing further opportunity to develop potentially effective drug therapies. The main goal of Phase II clinical trials is to identify a promising treatment to warrant a formal confirmatory Phase III trial. In this talk, we propose an optimal two-stage Bayesian adaptive design with multiple treatment arms to identify an effective treatment. Specifically, we use a Bayesian hierarchical model to improve operating characteristics of estimating the treatment effect in multiple treatment arms. Our approach provides an data-driven threshold after the first stage, achieving the maximum statistical power with controlling familywise type I error at desired level. Moreover, grid-search algorithm is developed to find a flexible optimal threshold. Our extensive simulation study shows the superior performance of the proposed approach over the existing trial design.

Recent Developments in Deep Learning/AI with Applications to Advance Pharmaceutical Research (IS-44)

Chair: Shuangge (Steven) Ma

Proposer: Huangdi (Denise) Yi, Shuangge (Steven) MA, Servier

Room: McHugh 101

Presenters: Huangdi Yi; Bino John; Siming Zheng; Durga V. Sritharan

Comparative Effectiveness Analysis of Lumpectomy and Mastectomy for Elderly Female Breast Cancer Patients: A Deep Learning-Based Big Data Analysis

Huangdi (denise) Yi, Department of Global Biometrics, Servier Pharmaceuticals

Abstract: Objectives: To evaluate the comparative effectiveness of treatments, a randomized clinical trial remains the gold standard but can be challenged by a high cost, a limited sample size, an inability to fully reflect the real world, and feasibility concerns. The objective is to showcase a big data approach that takes advantage of large electronic medical record (EMR) data to emulate clinical trials. To overcome the limitations of regression analysis, a deep learning-based analysis pipeline was developed. Study Design and Setting: Lumpectomy (breast-conserving surgery) and mastectomy are the two most commonly used surgical procedures for early-stage female breast cancer patients. An emulation trial was designed using the Surveillance, Epidemiology, and End Results (SEER)-Medicare data to evaluate their relative effectiveness in overall survival. The analysis pipeline consisted of a propensity score step, a weighted survival analysis step, and a bootstrap inference step. Results: A total of 65,997 subjects were enrolled in the emulated trial, with 50,704 and 15,293 in the lumpectomy and mastectomy arms, respectively. The two surgery procedures had comparable effects in terms of overall survival (survival year change = 0.08, 95% confidence interval (CI): -0.08, 0.25) for the elderly SEER-Medicare early-stage female breast cancer patients. Conclusion: This study demonstrated the power of "mining large EMR data + deep learning-based analysis," and the proposed analysis strategy and technique can be potentially broadly applicable. It provided convincing evidence of the comparative effectiveness of lumpectomy and mastectomy.

Emerging Opportunities to Accelerate Drug Discovery Using AI

Bino John, Boehringer Ingelheim

Abstract: This talk will focus on sharing next-generation ideas on how AI practitioners can help accelerate drug discovery. I will begin with the introduction of an overarching novel concept of enabling the concept of a digital patient using omics. This will be further followed by explaining additional avenues where AI can make an impact, spanning target to disease. Additional examples centered on the therapeutic modality that can help maximize efficacy, selectivity, safety, and clinical translation will be discussed.

Deep Domain Generalization

Siming Zheng, Yale

Abstract: We propose a domain-specific regression approach for domain generalization, taking into account possible heterogeneity among the datasets from different sources. In the proposed model, the domain-specific features are characterized through linear functionals of the marginal source distributions. The predictors are combined with the domain-specific linear functionals as inputs in the model. Using the source data sets, we estimate the domain-index function and the regression function nonparametrically based on neural network approximation. The resulting estimated domain-specific regression function can be used for prediction when future unlabelled data from a target domain arrives. The proposed method is shown to be consistent in the sense the cross-domain prediction error of the estimated domain-specific predictive function converges to that of the underlying true domain-specific predictive function. We also establish the convergence rates under suitable conditions. We demonstrate the performance of the proposed method through numerical experiments with simulated and real data. The numerical results show that, our method outperforms the naive pooling method and the nearest-domain-prediction method and our method is robust to model assumptions.

Deriving Imaging Biomarkers for Non-Small Cell Lung Cancer

Durga V. Sritharan, Yale School of Medicine

Abstract: Despite recent therapeutic advances, non-small cell lung cancer (NSCLC) represents a significant public health concern with 5-year overall survival ranging from 40-50% for early stage disease and 15-20%

for locally advanced disease respectively. Risk stratification of NSCLC patients is clinically challenging and there exist wide variations in outcomes among similarly staged patients. Quantitative imaging analysis has proven prognostic in a number of diseases and represents a potential low-cost biomarker to estimate cancer outcomes. In this talk, we will discuss our work investigating the utility of deep learning derived imaging-based biomarkers for improving prognostication for early stage and locally advanced NSCLC.

Innovations in Data Analysis: From Neural Networks to Bayesian Frameworks (IS-62)

Chair: Buddika Peiris

Proposer: Jian Zou and Buddika Peiris, Worcester Polytechnic Institute

Room: McHugh 201

Presenters: Bal Nandram (WPI); Buddika Peiris (WPI); Nadeesha Jayaweera (WPI); Tharindu De Alwis (WPI)

A Robust Bayesian Analysis of A Non-Probability Sample

Balgobin Nandram, Worcester Polytechnic Institute

Abstract: We stratify a finite population, and allocate a non-probability sample (nps) and a much smaller probability sample (ps) to the strata. The ps is mainly used to obtain sub-population sizes and total covariates because non-sampled covariates are not observed in the nps. We construct robust models to analyze the stratum data of the nps, mainly based on the Scott Smith model (without covariates) with spatial effects among the strata, and Bayesian pre dictive inference is done for small areas, which cut across the strata. We have an illustrative example of body mass index data for eight counties (small areas) from California. Keywords: Gibbs sampler, Mass imputation, Population model, Predictive inference, Sam ple model, Scott-Smith model, Stratification, Surrogate samples.

Restricted Inference in Circular-Linear and Linear-Circular Regression

Buddika Peiris, Worcester Polytechnic Institute

Abstract: In this work, we investigate restricted inference on two types of circular regression, called circular-linear and linear-circular. Our aim in this paper is to propose an alternative method which is necessary to apply where one observes a weak association between circular dependent and linear predictor variables, or between linear dependent and circular predictor variables, having clear knowledge about the sign of slope. We illustrate that restricted inference is particularly useful for those circular regressions, which is due to weak association. Comparison between our proposed restricted inference and the unrestricted inference are given by using two examples, one from ecological study and the other from environmental study. Keywords: Air quality, Amplitude of tide, Circular data, Slope parameter

A Bayesian Online Spatio-Temporal Detection Framework with Likelihood Weight Smoothing for Disease Surveillance

I.m.l. Nadeesha Jayaweera, Worcester Polytechnic Institute

Abstract: The dynamic nature of disease transmission, influenced by factors like population density, poses a significant challenge to accurate prediction. The study introduces a novel approach, integrating likelihood weighting into Integrated Nested Laplace Approximation (INLA) based on population density, to predict disease surveillance data through spatio-temporal Bayesian methodology. For non-stationary outbreak time series online prediction, our approach prioritizes accounting for more recent information with calibrated discounting on old information through weight adjustment on their likelihood. Empirical analysis on real COVID-19 daily case count data in Massachusetts counties demonstrates the effectiveness of our approach, showing improved prediction accuracy compared to existing methods. Our INLA-based method with weighted smoothing presents a promising avenue for enhancing infectious disease forecasting models, with potential applications in public health decision-making and resource allocation.

Stacking Based Neural Network for Nonlinear Time Series Analysis

Tharindu De Alwis, Worcester Polytechnic Institute (WPI)

Abstract: Stacked generalization is a commonly used technique for improving predictive accuracy by com-

bining less expressive models using a high-level model. This paper introduces a stacked generalization scheme specifically designed for nonlinear time series models. Instead of selecting a single model using traditional model selection criteria, our approach stacks several nonlinear time series models from different classes and proposes a new generalization algorithm that minimizes prediction error. To achieve this, we utilize a feed-forward artificial neural network (FANN) model to generalize existing nonlinear time series models by stacking them. Network parameters are estimated using a backpropagation algorithm. We validate the proposed method using simulated examples and a real data application. The results demonstrate that our proposed stacked FANN model achieves a lower error and improves forecast accuracy compared to previous nonlinear time series models, resulting in a better ft to the original time series data. Keywords: Stacked generalization, Cross-validation, Time series, Feed-forward artificial neural network (FANN), and Backpropagation algorithm.

Student Paper 2 (S2)

Chair: Lulu Kang

Organizer: Neil Spencer

Room: McHugh 306

A Bayesian Record Linkage Approach That Adjusts for Variables in One File

Gauri Kamat, Brown University

Co-authors: Mingyang Shan, Roee Gutman

Abstract: In many healthcare and social science applications, information about units is dispersed across multiple data files. Linking records across files is necessary to estimate associations between variables exclusive to each of the files. Common record linkage algorithms only rely on similarities between linking variables that appear in all the files. Moreover, analysis of linked files often ignores errors that may arise from incorrect or missed links. Bayesian record linking methods allow for natural propagation of linkage error, by jointly sampling the linkage structure and the model parameters. We extend an existing Bayesian record linkage method to integrate associations between variables exclusive to each file being linked. We show analytically, and using simulations, that the proposed method improves the linking process, and results in accurate inferences. We apply the method to link Meals on Wheels recipients to Medicare Enrollment records.

A Partially Randomized Patient Preference, Sequential, Multiple-Assignment, Randomized Trial Design Analyzed via Weighted and Replicated Frequentist and Bayesian Methods

Marianthie Wank, University of Michigan

Co-authors: Roy N. Tamura, Kelley M. Kidwell, Thomas M. Braun, Sarah Medley

Abstract: Results from randomized control trials (RCTs) may not be representative when individuals refuse to be randomized or are excluded for having a preference for which treatment they receive. If trial designs do not allow for participant treatment preferences, trials can suffer in accrual, adherence, retention, and external validity of results. Thus, there is interest surrounding clinical trial designs that incorporate participant treatment preferences. We propose a Partially Randomized, Patient Preference, Sequential, Multiple Assignment, Randomized Trial (PRPP-SMART) which combines a Partially Randomized, Patient Preference (PRPP) design with a Sequential, Multiple Assignment, Randomized Trial (SMART) design. This novel PRPP-SMART design is a multi-stage clinical trial design where, at each stage, participants either receive their preferred treatment, or if they do not have a preferred treatment, they are randomized. This paper focuses on the clinical trial design for PRPP-SMARTs and the development of Bayesian and frequentist weighted and replicated regression models (WRRMs) to analyze data from such trials. We propose a two-stage PRPP-SMART with binary end of stage outcomes and estimate the embedded dynamic treatment regimes (DTRs). Our WRRMs use data from both randomized and non-randomized participants for efficient estimation of the DTR effects. We compare our method to a more traditional PRPP analysis which only considers participants randomized to treatment. Our Bayesian and frequentist methods produce more efficient DTR estimates with negligible bias despite the inclusion of non-randomized participants in the analysis. The proposed PRPP-SMART design and analytic method is a promising approach to incorporate participant treatment preferences into clinical trial design.

Bayesian Semi-Supervised Inference via A Debiased Modeling Approach

Gözde Sert, Texas A&M University

Co-authors: Abhishek Chakrabortty and Anirban Bhattacharya

Abstract: Inference in semi-supervised (SS) settings has received a great amount of attention in recent years due to increased relevance in modern big-data problems. In a typical SS setting, there is a much larger sized unlabeled data containing only observations for predictors, in addition to a moderately sized labeled data involving observations for both an outcome and a set of predictors. Such data arises naturally from settings where the outcome, unlike the predictors, is costly to obtain. One of the primary statistical objectives in SS settings is to explore whether parameter estimation can be improved by exploiting the unlabeled data. This article proposes a novel Bayesian approach to SS inference for the population mean estimation problem. The proposed approach provides improved and optimal estimators both in terms of estimation efficiency as well as inference. The method itself has several interesting artifacts. The central idea behind our method is to model certain summary statistics of the data rather than specifying a probability model for the entire raw data itself. Specifying appropriate summary statistics crucially relies on a debiased representation of the population mean in terms of nuisance parameters. Combined with careful usage of sample splitting, our debiasing approach mitigates the effect of bias due to slow rates or misspecification of the nuisance parameter from the posterior of the final parameter of interest. We establish concrete theoretical results, via Bernstein–von Mises theorems, validating all our claims and further supporting them through extensive numerical studies. To the best of our knowledge, this is the first work in Bayesian inference for SS settings. We also believe that the central idea of this article will be more broadly applicable to Bayesian semi-parametric inference.

Asymptotic Bayes Optiamlity for Sparse Count Data

Sayantan Paul, Indian Statistical Institute, Kolkata, India

Abstract: Consider a situation of analyzing high-dimensional count data containing an excess amount of zeroes and small non-zero counts. Under the assumption that the observations are modeled by a Poisson distribution, we are interested in simultaneous testing of the means of those observations. In this work, we study some optimal properties (in terms of Bayes risk) of multiple-testing rules induced by both two-groups and a general class of one-group shrinkage priors. The class of one-group priors was proposed by Polson and Scott (2010) and studied by Ghosh et al. (2016) in the context of the normal means model and by Tang et al. (2018) for the linear regression model. Here, first, we model each mean by a two-groups prior, and under the assumption of $0-1$ loss function, we obtain an expression for the optimal Bayes risk under some assumption similar to Bogdan et al. (2011). Next, same as Ghosh et al. (2016), we assume a two-groups mixture model for the data and model each mean parameter by the broad class of one-group priors to study the Bayes risk induced by the chosen class of priors. We have been able to show that, when the underlying sparsity pattern is known, under some proposed assumptions, the Bayes risk corresponding to the broad class of priors attains the optimal Bayes risk exactly. In other words, it ensures the decision rule is Asymptotically Bayes optimal under sparsity (ABOS). When this sparsity is unknown, motivated by Yano et al. (2021), we use an empirical Bayes estimate of the global shrinkage parameter. In this case, also, we show that the modified decision rule is ABOS, too. In this way, as an alternative solution for two-groups prior, we propose a broad class of global-local priors having similar optimal properties in terms of Bayes risk to quasi-sparse count data.

Combining BART and Principal Stratification to Estimate The Effect of Intermediate on Primary Outcomes with Application to Estimating The Effect of Family Planning on Employment in Sub-Saharan Africa.

Lucas Godoy Garraza, Department of Biostatistics and Epidemiology, University of Massachusetts Amherst

Co-authors: Leontine Alkema (Dep. of Biostatistics and Epidemiology, University of Massachusetts Amherst); Ilene S Speizer (Dep. of Maternal and Child Health, University of North Carolina at Chapel Hill Gillings)

Abstract: Motivation: There is interest in learning about the causal effect of family planning (FP) on empowerment related outcomes. Experimental data related to this question is available from trials in which FP programs increase access to FP. While program assignment is unconfounded, FP uptake and subsequent empowerment may share common causes. Methods: We use principal stratification to estimate the causal

effect of an intermediate FP outcome on a primary outcome of interest, among women affected by a FP program. Within strata defined by the potential reaction to the program, FP uptake is unconfounded. To minimize the need for parametric assumptions, we propose to use Bayesian Additive Regression Trees (BART) for modeling stratum membership and outcomes of interest. We refer to the combined approach as Prince BART. We evaluate Prince BART a simulation study. Results: We use prince BART to assess causal effect of modern contraceptive use on employment in six cities in Nigeria, based on quasi-experimental data from a FP program trial during the first half of the 2010s. We show that findings differ between Prince BART and approaches based on parametric assumptions.

Parallel Session 3 | 04:00 PM - 05:40 PM, May 22

Advances in Statistical Machine Learning with Innovative Applications (IS-9)

Chair: Xingche Guo

Proposer: Xingche Guo, Department of Biostatistics, Columbia University

Room: McHugh 101

Presenters: Dan Nettleton; Annie J Lee; Shan Yu; Bo Shen

A Random Forest Prediction Interval with Coverage Guarantees

Dan Nettleton, Iowa State University

Abstract: We consider the problem of quantifying uncertainty associated with a random forest prediction. In particular, we present a prediction interval for an unknown response value that is constructed from a random forest prediction and the empirical distribution of the random forest's out-of-bag prediction errors. The prediction interval is computationally inexpensive because it can be computed from the fit of a single random forest and its byproducts. We show that the interval has good empirical performance with respect to width and coverage across many example datasets. Furthermore, we use the concept of stability to provide non-asymptotic lower and upper bounds on interval coverage. The result is a straightforward approach for generating prediction intervals with confidence from the standard random forest algorithm heavily used in practice.

A Semi-Parametric Approach for Longitudinal Outcome-Guided Disease Subtyping Using High-Dimensional Omics Data

Annie J Lee, Columbia University

Abstract: Primary challenges to treating and preventing neurodegenerative diseases include extensive heterogeneity in the clinicopathologic state of older individuals, suggesting the presence of subgroups of individuals who share certain biological features but differential responses to disease risk factors. High-throughput sequencing and conventional unsupervised clustering methods have been employed to identify subgroups of individuals who share similar patterns of genomic features to define disease subtypes. The resulting clusters, however, may not capture clinically meaningful disease subtypes because identified clusters may not relate to clinical outcomes and confounders (e.g., demographic factors) can dominate the clustering procedure. To identify disease subtypes guided by a clinical outcome, existing methods, such as supervised clustering, use mixture models that focus on cross-sectional clinical outcomes and covariates. In this talk, we propose a novel latent mixture model that incorporates longitudinal clinical outcomes and time-varying covariates to identify outcome-guided disease subtypes from high-dimensional omics data. Our approach defines clinically meaningful subtypes based on the association of risk factors and disease clinical outcome of interest and incorporates genetic pathway information to regularize variable selection when predicting subgroup membership. The proposed method will be applied to investigate health disparities in Alzheimer's disease among elderly non-Hispanic Whites and Hispanics by leveraging transcriptomic profiles and longitudinal clinical data from two studies of aging and dementia. Through the identification of clinically relevant disease subtypes, we will investigate the clinicopathologic and neurobiological relevance of these subtypes and pinpoint genes or pathways that are unique to one ethnic group or are common among both non-Hispanic Whites and Hispanics. Through a special construction, we will also test genes by cardiovascular risk factor interaction. This work is crucial for the design of therapeutics and trials that focus on precise molecular targets that can be better tailored to individuals or certain ethnic groups.

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

Shan Yu, University of Virginia

Abstract: Motivated by recent work analyzing data in biomedical imaging studies, we consider a class of functional regression models for imaging responses. We introduce a novel nonparametric distributed (NPD) learning framework that utilizes multivariate spline smoothing over a triangulation of domain. The proposed NPD estimation algorithm features a scalable and communication-efficient implementation scheme to achieve near-linear speedup. Asymptotic confidence intervals and data-driven simultaneous confidence corridors (SCCs) for the coefficient functions are constructed. Our method can simultaneously estimate and make inferences of the coefficient functions while incorporating the spatial heterogeneity. In addition, we provide rigorous theoretical support for the NPD

estimation and inference framework. Specifically, we demonstrate that the NPD-based spline estimators are asymptotic normal and have the same convergence rate as the global spline estimators obtained using the entire dataset. Monte Carlo simulation studies are conducted to examine the finite-sample performance of the proposed method. The proposed method is applied to the spatially normalized Positron Emission Tomography (PET) data of Alzheimer's Disease Neuroimaging Initiative (ADNI).

Bayesian Optimization for Design Problems in Manufacturing

Bo Shen, Department of Mechanical and Industrial Engineering, New Jersey Institute of Technology

Abstract: In manufacturing, optimizing design parameters is critical for achieving superior product performance, cost-efficiency, and quality. However, the complexity of design spaces and the expensive nature of experimentation pose significant challenges. Bayesian optimization emerges as a potent methodology to tackle these challenges by efficiently exploring the design space and identifying optimal solutions.

Recent Statistical Methods and Machine Learning Algorithms for Electronic Health Records (IS-25)

Chair: JooChul Lee

Proposer: JooChul Lee, University of Pennsylvania

Room: McHugh 206

Presenters: Weidong Ma; Huan He; Sarah E. Hegarty; Joochul Lee

A Semiparametric Method for Addressing Under-Diagnosis Using Electronic Health Record Data

Weidong Ma, Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania Perelman School of Medicine

Abstract: Under-diagnosis, which occurs when patients live with a disease condition without receiving a diagnosis, prevents patients from obtaining suitable treatments and preventive strategies. Electronic Health Records (EHRs) contain a wealth of patient information and offer a unique opportunity to identify under-diagnosis, as diagnosed and under-diagnosed patients may exhibit similarities in their EHR profiles, which differ from those of disease-free patients. However, this opportunity to date has not been fully exploited due to the "positive-unlabeled" structure of EHR data, where under-diagnosed patients are mixed together with a large number of disease-free patients. To address this challenge, we develop a novel statistical approach based on importance weighting method to enable unbiased assessment of the risk that a patient has the disease condition, where EHR data is supplemented with a small number of additional disease labels acquired through targeted screening for patients who have not received a diagnosis. The performance of the proposed method is studied via characterization of asymptotic properties and extensive simulation studies. We apply our approach to Penn Medicine EHRs to identify patients under-diagnosed with non-alcoholic steatohepatitis (NASH).

A Flexible Generative Model for Heterogeneous Tabular EHR with Missing Modality

Huan He, University of Pennsylvania

Abstract: Realistic synthetic electronic health records (EHRs) can be leveraged to accelerate methodological developments for research purposes while mitigating privacy concerns associated with data sharing. However, the training of Generative Adversarial Networks remains challenging, often resulting in issues like mode col- lapse. While diffusion models have demonstrated progress in generating quality synthetic samples for tabular EHRs given ample denoising steps, their performance wanes when confronted with missing modalities in heterogeneous tabular EHRs data. For example, some EHRs contain solely static measurements, and some contain only contain temporal measurements, or a blend of both data types. To bridge this gap, we introduce FLEXGEN-EHR– a versatile diffusion model tailored for heterogeneous tabular EHRs, equipped with the capability of handling missing modalities in an integrative learning framework. We define an optimal transport module to align and accentuate the common feature space of heterogeneity of EHRs. We empirically show that our model consistently outperforms existing state-of-the-art synthetic EHR generation methods both in fidelity by up to 3.10% and utility by up to 7.16%. Additionally, we show that our method can be suc-

cessfully used in privacy-sensitive settings, where the original patient-level data cannot be shared.

Algorithm Fairness in Healthcare: Error-Rate Fairness Definitions in The Presence of Risk Distribution Differences

Sarah E. Hegarty, University of Pennsylvania

Abstract: Algorithms range in complexity from simple rule-based workflows to black box machine learning models. As they can be readily implemented on massive scales, any differences in performance can lead to systematic, differential impacts. This concern has prompted a flurry of activity to define and evaluate algorithm fairness statistically. Many of these approaches have centered on error-rate metrics such as true positive rate, true negative rate or positive predictive value. While holding some intuitive appeal, forcing models to satisfy parity in such error-rate metrics results in an overall loss of predictive accuracy. Moreover, parity in these metrics cannot be satisfied simultaneously except in trivial settings. Little guidance has been offered for which aspects of fairness to prioritize or how to navigate the fairness-accuracy tradeoff. In the current work, we expose two primary mechanisms by which popular error-rate based definitions of fairness, such as equal opportunity, can be violated: differences in the underlying risk distributions and differential calibration bias. Through a series of numerical studies, we demonstrate how such metrics are expected to differ between subgroups in the presence of differences in underlying risk distribution and/or subgroup-depended calibration bias. Moreover, we propose an adjustment to the popular error-rate definitions that control for potential confounding from the underlying risk distributions. We demonstrate the performance of the adjusted metrics through numerical studies and a real-word data application to a risk-prediction model in use in a large urban health system.

Towards Optimal Model Evaluation: Enhancing Active Testing with Actively Improved Estimators

Joochul Lee, University of Pennsylvania

Abstract: A common challenge for validating a risk prediction model using electronic health record (EHR) data is that labels for the predicted outcome are not directly available. Towards efficient and unbiased model validation, we study optimal sampling designs for efficiently labeling an informative subset of patients in an EHR cohort. Given a pre-specified number of outcome labels, our design aims to minimize the asymptotic variance of an improved inverse probability weighted estimator for predictive accuracy metrics. Implementation of the optimal sampling requires accurate risk estimates and the predictive accuracy metric of interest. We therefore propose to implement sampling in two steps. First a portion of the target number of labels is acquired by applying entropy sampling to a random subset of the cohort. These initial labels are then used to calibrate risk estimates and obtain an initial estimate of the predictive accuracy metric, which are then used to inform optimal sampling of the remaining target number of labels. The final estimate of the predictive accuracy metrics is obtained by applying the proposed estimator to the full cohort and all acquired labels pooled together. Furthermore, we address this issue by extending existing work on "Active Testing" (AT) methods which are designed to sequentially sample and label data for the evaluation of pre-trained models. Application to a real EHR dataset indicate superior efficiency of the proposed sampling design and the proposed estimator.

Multiplicity Control in Innovative Drug Development Applications (IS-26)

Chair: Lei Nie

Proposer: Lei Nie, The US FDA

Room: McHugh 202

Presenters: Frank Bretz; Werner Brannath; Dong Xi; Discussant George Kordzakhia

Optimal Test Procedures for Multiple Hypotheses Controlling The Familywise Expected Loss

Frank Bretz, Novartis

Abstract: We consider the problem of testing multiple null hypotheses, where a decision to reject or retain must be made for each one and embedding incorrect decisions into a real life context may inflict different losses. We argue that traditional methods controlling the Type I error rate may be too restrictive in this situation and that the standard familywise error rate may not be appropriate. Using a decision-theoretic

approach, we define suitable loss functions for a given decision rule, where incorrect decisions can be treated unequally by assigning different loss values. Taking expectation with respect to the sampling distribution of the data allows us to control the familywise expected loss instead of the conventional familywise error rate. Different loss functions can be adopted, and we search for decision rules that satisfy certain optimality criteria within a broad class of decision rules for which the expected loss is bounded by a fixed threshold under any parameter configuration. We illustrate the methods with the problem of establishing efficacy of a new medicinal treatment in non-overlapping subgroups of patients. Reference: Maurer, Willi, Frank Bretz, and Xiaolei Xun (2023) Optimal test procedures for multiple hypotheses controlling the familywise expected loss (with Discussion). Biometrics 79:2781–2793.

Control of Essential Type I Errors

Werner Brannath, University of Bremen, Faculty 3 - Mathematics and Computer Science, Institute for Statistics & KKSB

Abstract: The talk will discuss alternatives to controlling the familywise error rate in multiple hypothesis clinical trials. The focus will be on concepts that control Type I error rates only insofar as they are relevant to patients outside and after the trial. With a focus on multi-population studies, the population-wise error rate (PWER; Brannath et al., 2022) will be introduced as an example and related to the familywise expected loss (FWEL; Maurer et al., 2022). In addition, for multi-arm and platform trials with the possibility to drop treatments mid-trial, it will be discussed how to account for a mid-trial reduction in post-trial risks when dropping a treatment. The solutions will be motivated by independent clinical trials that do not require multiplicity adjustment. The talk will conclude with a discussion and outlook on further questions and future research on the control of essential type I error rates. References - Brannath, W., Hillner, C., Kornelius, R. (2023). The population-wise error rate for clinical trials with overlapping populations. Statistical Methods in Medical Research, 32(2):334-352 - Maurer, W., Bretz, F., and Xun, X. (2023). Optimal test procedures for multiple hypotheses controlling the familywise expected loss. Biometrics 79(4):2781-2793 - Brannath, W. (2023). Discussion on "Optimal test procedures for multiple hypotheses controlling the familywise expected loss" by Willi Maurer, Frank Bretz, and Xiaolei Xun. Biometrics 79(4):2806-2810.

New Strategies for Confirmatory Testing of Secondary Hypotheses on Combined Data From Multiple Trials

Dong Xi, Gilead Sciences

Abstract: Pivotal evidence of efficacy of a new drug is typically generated by (at least) two clinical trials which independently provide statistically significant and mutually corroborating evidence of efficacy based on a primary endpoint. In this situation, showing drug effects on clinically important secondary objectives can be demanding in terms of sample size requirements. Statistically efficient methods to power for such endpoints while controlling the Type I error are needed. We review existing strategies for establishing claims on important but sample size–intense secondary endpoints. We present new strategies based on combined data from two independent, identically designed and concurrent trials, controlling the Type I error at the submission level. We explain the methodology and provide three case studies. Different strategies have been used for establishing secondary claims. One new strategy, involving a protocol planned analysis of combined data across trials, and controlling the Type I error at the submission level, is particularly efficient. It has already been successfully used in support of label claims. Regulatory views on this strategy differ. Inference on combined data across trials is a useful approach for generating pivotal evidence of efficacy for important but sample size–intense secondary endpoints. It requires careful preparation and regulatory discussion.

Innovative Statistical Methods for Longitudinal and Survival Data Analyses (IS-36)

Chair: Panpan Zhang

Proposer: Panpan Zhang, Vanderbilt University Medical Center

Room: McHugh 301

Presenters: Hongyuan Cao; Shou-en Lu; Jing Qian; Jun Yan

Kernel Meets Sieve: Transformed Hazards Models with Sparse Longitudinal Covariates

Hongyuan Cao, Florida State University

Abstract: We study the transformed hazards model with time-dependent covariates observed intermittently for the censored outcome. Existing work assumes the availability of the whole trajectory of the time-dependent covariates, which is unrealistic. We propose combining kernel-weighted log-likelihood and sieve maximum log-likelihood estimation to conduct statistical inference. The method is robust and easy to implement. We establish the asymptotic properties of the proposed estimator and contribute to a rigorous theoretical framework for general kernel-weighted sieve M-estimators. Numerical studies corroborate our theoretical results and show that the proposed method performs favorably over competing methods. The analysis of a data set from a COVID-19 study in Wuhan identifies clinical predictors that otherwise cannot be obtained using existing methods.

Fixed and Random Effect Selections in Generalized Linear Mixed Models

Shou-en Lu, Rutgers School of Public Health

Abstract: Generalized linear mixed models are commonly used to describe relationships between correlated responses and covariates in medical research. In this paper, we propose a simple and easily implementable regularized estimation approach to select both fixed and random effects in generalized linear mixed model. Specifically, we propose to construct and optimize the objective functions using the confidence distributions of model parameters, as opposed to using the observed data likelihood functions, to perform effect selections. Two estimation methods are developed. The first one is to use the joint confidence distribution of model parameters to perform simultaneous fixed and random effect selections. The second method is to use the marginal confidence distributions of model parameters to perform the selections of fixed and random effects separately. With a proper choice of regularization parameters in the adaptive LASSO framework, we show the consistency and oracle properties of the proposed regularized estimators. Simulation studies have been conducted to assess the performance of the proposed estimators and demonstrate computational efficiency. Our method has also been applied to two longitudinal cancer studies to identify demographic and clinical factors associated with patient health outcomes after cancer therapies.

Estimation and Regression Analysis with Sequentially Truncated Survival Data

Jing Qian, University of Massachusetts Amherst

Abstract: In observational cohort studies with complex sampling schemes, truncation arises when the time to event of interest is observed only when it falls below or exceeds another random time, i.e., the truncation time. In more complex settings, observation may require a particular ordering of event times; we refer to this as sequential truncation. We propose nonparametric and semiparametric maximum likelihood estimators for the distribution of the event time of interest in the presence of sequential truncation, under two truncation models. We develop methods for regression modeling in this complex setting using the tool of pseudo-observations. We evaluate our approach in simulation studies and in application to an Alzheimer's cohort study.

Recurrent Events Modeling Based on A Reflected Brownian Motion with Application to Hypoglycemia

Jun Yan, University of Connecticut

Abstract: Patients with type 2 diabetes need to closely monitor blood sugar levels as their routine diabetes self-management. Although many treatment agents aim to tightly control blood sugar, hypoglycemia often stands as an adverse event. In practice, pa- tients can observe hypoglycemic events more easily than hyperglycemic events due to the perception of neurogenic symptoms. We propose to model each patient's observed hypoglycemic event as a lower-boundary crossing event for a reflected Brownian motion with an upper reflection barrier. The lower-boundary is set by clinical standards. To capture patient heterogeneity and within-patient dependence, covariates and a patient level frailty are incorporated into the volatility and the upper reflection barrier. This framework provides quantification for the underlying glucose level variability, patients heterogeneity, and risk factors' impact on glucose. We make inferences based on a Bayesian framework using Markov chain Monte Carlo. Two model comparison crite- ria, the Deviance Information Criterion and the Logarithm of the Pseudo-Marginal Likelihood, are used for model selection. The methodology is validated in simulation studies. In analyzing a dataset from the diabetic patients in the DURABLE trial, our model provides adequate fit, generates data similar to the observed data, and offers insights that could be missed by other models.

Advanced Methods for Analyzing Time Series Data (IS-47)

Chair: Yao Zheng

Proposer: Nalini Ravishanker, University of Connecticut

Room: McHugh 305

Presenters: Sreeram Anantharaman; Shibo Li; Zhaoyuan Li; Patrick Toman

Modeling Multiple Synchronized Irregularly-Spaced High Frequency Financial Time Series

Sreeram Anantharaman, University of Connecticut

Abstract: Multivariate volatility modeling in intraday scenarios enables to capture the complex interrelationships and dependencies among multiple asset price movements over short time intervals, facilitating more accurate risk assessment and portfolio management strategies in high-frequency trading environments. However, intraday data poses unique challenges due to its irregular and high-frequency nature necessitating the development of new model classes to effectively analyze multivariate volatility. Specifically, these models must incorporate the irregular gaps between successive transactional events to better understand market microstructure. We propose a hierarchical irregular basic multivariate stochastic volatility autoregressive conditional duration (IR-BMSV-ACD) model that integrates autoregressive conditional duration (ACD) model for modeling the inter-transaction gaps. This integration enables the estimation of the future volatility of log-returns, providing a more comprehensive understanding of the future market dynamics. Our analysis is conducted within a Bayesian framework using the Hamiltonian Monte Carlo (HMC) algorithm with No-U-turn sampler (NUTS) in R using the cmdstanr package. Fitting and forecasting are performed through Monte Carlo averages based on posterior samples. We demonstrate this methodology through simulation studies and real-data analysis of intra-day prices of health stocks traded on the NYSE at the microsecond level. Employing the refresh time sampling technique, we synchronize the transactions occurring at different times for each stock. Subsequently, we compute the synchronized log-returns and gaps for the stocks, which are then utilized for modeling purposes.

Tensor Autoregression via Candecomp/Parafac Decomposition

Shibo Li, University of Connecticut

Abstract: As tensor-valued time series commonly appear in various fields, efficient methods to achieve dimension reduction in modeling high-dimensional tensor-valued time series are highly demanded. We propose a new model to realize low-rank and sparsity assumptions in tensor autoregression. The CANDECOMP/PARAFAC (CP) decomposition is applied to induce a multi-dimensional low-rank structure, which simultaneously decomposes the transition tensor into several blocks along different directions of the tensor. The selection of important features in the decomposition components is achieved by a truncating method. We derive the non-asymptotic property for the proposed estimator and design an algorithm for the proposed model. Simulation studies verify the theoretical results and demonstrate the advantage of our algorithm in computational time. An application to air pollution data demonstrates interesting structural interpretations unveiled by the proposed model.

Online Change-Point Detection and El Niño Prediction

Zhaoyuan Li, The Chinese University of Hong Kong, Shenzhen

Abstract: The problem of change point detection in the correlation structure of streaming high-dimensional data is explored, with minimum assumptions posed on the underlying data distribution and correlation structure. Sparse and dense setting are considered. Both window-limited and Shewhart-type test statistics are proposed. A novel method for threshold selection is designed based on sign-flip permutation. In addition, two enhancement techniques, synthetic minority oversampling technique (SMOTE) and knockoff, are combined with window-limited test statistics to tackle the instability in detection due to small sample sizes. Theoretical evaluations of these proposed methods are conducted in terms of average run length and expected detection delay. Numerical studies are conducted to examine the finite sample performances of the proposed methods. We concentrate on the changes in correlations and show that well before an El Nino episode the correlations tend to increase first and then decrease sharply. We use this robust observation to forecast El Niño development in advance. This is the first time that an online change point detection procedure is used to predict El Niño events. Our method yields hit rates above 0.85 and false-alarm rates smaller than 0.1, based on high-quality observational data.

Gaussian Process State Space Models with Applications to Iot Time Series

Patrick Toman, Hartford Steam Boiler

Abstract: The proliferation of "Internet of Things" (IoT) sensor technology has led to the creation of large time series datasets with complex dependence structures that cannot be readily modeled by conventional, linear time se ries models such as Bayesian structural time series (BSTS). As such, one must turn to more sophisticated methods that can grapple with the inher ent challenges of IoT time series. One potential solution is the Gaussian process state space model (GPSSM) which allows for flexible, probabilistic modeling of complex temporal data by placing a Gaussian process (GP) prior over the latent state transition function. By leveraging modern vari ational inference techniques, one can readily fit GPSSMs to a large number of time series in relatively short amounts of time. In this presentation, we first give an overview of GPSSMs with an emphasis on the usage of varia tional inference for training. Subsequently, we demonstrate the utility of GPSSMs by applying them to a time series intervention analysis problem for IoT temperature sensors.

Statistics and Computation in The Era of AI (IS-49)

Chair: Donghui Yan

Proposer: Donghui Yan, University of Massachusetts Dartmouth

Room: McHugh 205

Presenters: Yang Feng; Brendan Mcveigh; Jason Dou; Yuchen Fama; Jin Cao

Robust Unsupervised Multi-Task and Transfer Learning on Gaussian Mixture Models

Yang Feng, New York University

Abstract: Unsupervised learning has been widely used in many real-world applications. One of the simplest and most important unsupervised learning models is the Gaussian mixture model (GMM). In this work, we study the multi-task learning problem on GMMs, which aims to leverage potentially similar GMM parameter structures among tasks to obtain improved learning performance compared to single-task learning. We propose a multi-task GMM learning procedure based on the EM algorithm that not only can effectively utilize unknown similarity between related tasks but is also robust against a fraction of outlier tasks from arbitrary distributions. The proposed procedure is shown to achieve minimax optimal rate of convergence for both parameter estimation error and the excess mis-clustering error, in a wide range of regimes. Moreover, we generalize our approach to tackle the problem of transfer learning for GMMs, where similar theoretical results are derived. Finally, we demonstrate the effectiveness of our methods through simulations and real data examples. To the best of our knowledge, this is the first work studying multi-task and transfer learning on GMMs with theoretical guarantees.

Rare-Event Sampling for Autonomous Driving System Evaluation

Brendan Mcveigh, Waymo

Abstract: As autonomous driving systems (ADS) begin to outperform human benchmarks, high-consequence driving events have become increasingly rare: after more than 7 million miles without a human behind the wheel, the Waymo Driver has demonstrated reported crash rates lower than human benchmarks, fewer than one incident per million miles (IPMM). Given the rarity of high-consequence events and the computational cost of high-fidelity ADS simulation, efficient sampling of such events has become increasingly critical to ADS evaluation. From a corpus of simulated ADS behavior over millions of driving log segments, we used learned embeddings to predict which segments would contain one or more driving events of interest. The target events varied in prevalence by several orders of magnitude. The predictions were used to construct a biasing distribution for importance sampling. Our approach produced improved estimates of several distinct driving events of interest in tandem, increased the statistical power of detecting relevant changes to ADS behavior, and was computationally performant at scale.

Towards Ai-Enabled Healthcare Through Learning Effective Representations Efficiently

Jason Dou, Harvard Innovation Labs

Abstract: Learning effective representations efficiently

plays a pivotal role in machine learning applications ranging from computer vision, and natural language processing, to healthcare, mobile sensing, and computational biology. Given a large amount of informative but also noisy data in various domains, one of the core quests for artificial intelligence is motivating the design and analysis of new representation learning methods for various applications. The research objectives of this talk are to understand and develop new representation learning models through the lens of optimization and measurement, apply representation learning frameworks for various applications, and invent new metrics to creatively evaluate representation learning outcomes. First, Adaptive Sampling with Reward and Deep Wasserstein Learning frameworks are presented to tackle the sampling and measurement challenges in ranking-based loss functions with applications in computer vision and single-cell biology. Second, I will present data-efficient learning methods for mobile sensing and cross-model single cell data through coreset. Next, I will present a case study using representation learning to understand lung cancer patients' treatment trajectories after immunotherapy. Lastly, three evaluation metrics based on the science of science, information theory, and causality respectively are presented for knowledge measurement in knowledge representations. Taken together, this talk outlines a highly comprehensive, impactful, and interdisciplinary approach to representation learning research, from the perspectives of models, applications, and metrics, with broader implications for artificial intelligence and machine learning more generally.

Uncertainty-Aware Llms and Uncertainty Quantification with Bayesian Computation

Yuchen Fama, Normal Computing

Abstract: In this talk, we will provide a brief overview of uncertainty quantification and Bayesian computation, and applications solving frontier problems in AI especially LLMs. We introduce posteriors - a new open source Python library from Normal Computing that provides native tools using PyTorch and its functional API, and example demonstrations on how it can robustify predictions and avoid catastrophic forgetting.

Leveraging Statistical Inference: Casuality, Multiple Testing, and Uncertainty Quantification (IS-66)

Chair: TBD

Proposer: Kun Chen, University of Connecticut

Room: McHugh 307

Presenters: Yihan Bao; Kaiwen Hou; Shane Sacco; Zheng Gao

Estimating Causal Effects for Binary Outcomes Using Per-Decision Inverse Probability Weighting

Yihan Bao, Yale University

Abstract: Micro-randomized trials are commonly conducted for optimizing mobile health interventions such as push notifications for behavior change. In analyzing such trials, causal excursion effects are often of primary interest, and their estimation typically involves inverse probability weighting (IPW). However, in a micro-randomized trial, additional treatments can often occur during the time window over which an outcome is defined, and this can greatly inflate the variance of the causal effect estimator because IPW would involve a product of numerous weights. To reduce variance and improve estimation efficiency, we propose a new estimator using a modified version of IPW, which we call "per-decision IPW". It is applicable when the outcome is binary and can be expressed as the maximum of a series of sub-outcomes defined over subintervals of time. We establish the estimator's consistency and asymptotic normality. Through simulation studies and real data applications, we demonstrate substantial efficiency improvement of the proposed estimator over existing estimators (relative efficiency up to 1.45 and sample size savings up to 31% in realistic settings). The new estimator can be used to improve the precision of primary and secondary analyses for micro-randomized trials with binary outcomes.

Geometries of Efficient Semiparametric Estimation: Universal Least Favorable Flows and Alignment in Tangent Sets

Kaiwen Hou, Columbia University

Abstract: This paper presents a novel method for efficient semiparametric estimation that enhances geometric awareness by combining continuous normalizing flows (CNFs) with parametric submodels. This combination improves upon the traditional Targeted

Maximum Likelihood Estimation (TMLE) by optimizing the Cramér-Rao lower bound, resulting in a universal least favorable model. By using CNFs to follow these least favorable directions, our approach pushes forward a baseline distribution to a data-driven distribution where maximum likelihood estimation can be performed effectively. A key aspect of the theoretical justification of efficiency is aligning the velocity field in the continuity equation of CNFs with the tangent vectors of a regular semiparametric model. This alignment is further elucidated through the lens of Wasserstein gradient flows, which impose structured geometric constraints related to the Cramér-Rao lower bound to ensure a trajectory that minimizes estimation variance. Preliminary experimental results underscore the superiority of our approach, consistently yielding lower mean-squared errors compared to traditional methods such as TMLE. These outcomes emphasize the potential of geometry-aware flows to significantly enhance semiparametric estimation.

An Active Learning Approach to Help Account for Prediction Uncertainty of Models Deployed in Healthcare

Shane J Sacco, University of Connecticut

Abstract: Predicting health outcomes utilizing statistical models is becoming a more commonplace notion in healthcare. However, all statistical models contain some degree of uncertainty in prediction. Many current modelling approaches involve creating thresholds to transform model outputs into binary decisions regarding risk (i.e., 1=high risk; 0=not high risk). These model outputs are typically derived from point estimates of risk and do not account for uncertainty (or the possible error) behind this estimate. Prediction intervals may be constructed to describe uncertainty in point estimates, and for those in which we are uncertain, we may be able to leverage active learning. Particularly, additional information may be integrated into the decision-making process to either accept or reject the level of risk indicated by point estimates. This presentation explores how multiple active learning methods may help account for uncertainty in predictive models via Monte Carlo simulation and two case studies in context of major adverse cardiac events and suicide attempts. We found that active learning may enhance model performance, specifically for point estimates in which we are uncertain.

Laws of Large Dimensions

Zheng Gao, Upstart Network, Inc.

Abstract: Motivated by genome-wide association screening studies (GWAS), we revisit the classical problem of high-dimensional marginal screenings of categorical variables. We discuss some recent results on the phase transitions in large-scale multiple testing, and characterize four new phase transitions in high-dimensional chi-square models. Remarkably, degrees of freedom in the chi-square distributions do not affect the boundaries in all four phase transitions. Several well-known procedures are shown to attain these boundaries. We then elucidate on the nature of signal sizes in association tests, by characterizing its relationship with marginal frequencies, odds ratio, and sample sizes in 2x2 contingency tables. This allows us to illustrate an interesting manifestation of the phase transition phenomena in GWAS. We also show, perhaps surprisingly, that given total sample sizes, balanced designs in such association studies rarely deliver optimal power.

Statistical Challenges in Cell and Gene Therapies (IS-67)

Chair: Fengjuan (Joan) Xuan

Proposer: Glen Laird, Vertex Pharmaceuticals

Room: McHugh 201

Presenters: Lixi Yu; Weidong Zhang; Yingtian Hu; Avery Isaac Mcintosh

Discussion about Statistical Analysis Methods for External Control Comparison

Lixi Yu, Sarepta Therapeutics

Abstract: Although randomized controlled trials (RCTs) are the gold standard to assess the efficacy and safety of a new treatment, they are not always plausible. Reasons such as ethical considerations, small sample sizes in rare diseases make it inevitable in certain scenarios to borrow information from external controls in order to estimate treatment efficacy and safety. We propose a series of steps to consider when deciding to adopt an external control strategy discuss different statistical methods and conclude with data results and a discussion of unresolved issues.

Landscape and Clinical Design Considerations in Cell and Gene Therapy Development.

Weidong Zhang, Sana Biotechnology, Inc.

Abstract: Advances in cell and gene engineering technologies have given rise to an exponential growth of development of cell and gene therapies (CGT) around the world. The potential benefits of CGT have been explored in a broad range of therapeutic areas including oncology, rare diseases, diabetes, cardiovascular and CNS. In this talk, a comprehensive overview of key concepts, technologies and CGT landscape will first be introduced. In addition, key issues and considerations about cellular kinetics, dose finding strategy, clinical designs and analysis will also be discussed. Case studies using real world examples from a few key players will be presented to highlight the challenges and how statisticians can help address them.

Missing Data for Binary Endpoint in Gene and Cell Therapy

Yingtian Hu, Vertex Pharmaceuticals Inc

Abstract: In the development of gene and cell therapies, a small uncontrolled single arm study has often been employed for registration purpose. With the limited sample size, how to handle missing data could have a critical impact on the success of the study, especially when the primary interest is a binary endpoint. Missing data due to treatment-related and nonrelated drop-out require different analytic strategies. Non-responder imputation, last observation carried forward (LOCF), simple imputation and multiple imputation are typical methods to handle missing data. In this talk, we will discuss these methods in the setting of binary endpoint in gene and cell therapy.

Development of Gene Therapies: Open Questions on Design and Analysis

Avery Mcintosh, Pfizer

Abstract: Gene and cell therapies represent the apex of complexity in biopharmaceutics. The design, analysis, and clinical development of these products is complex and unique among drug development modalities. This talk will give a brief overview of gene therapies as a pharmaceutical drug class, with a focus on outstanding design and statistical issues in the development of these products, such as dose finding, long-term follow-up, and opportunities for adaptivity.

Student Paper 3 (S3)

Chair: Victor Hugo

Organizer: Neil Spencer

Room: McHugh 306

Spectrum-Aware Adjustment: A New Debiasing Framework with Applications to Principal Component Regression

Yufan Li, Harvard University

Co-authors: Pragya Sur

Abstract: We introduce a new debiasing framework for high-dimensional linear regression that bypasses the restrictions on covariate distributions imposed by modern debiasing technology. We study the prevalent setting where the number of features and samples are both large and comparable. In this context, state-of-the-art debiasing technology uses a degrees-of-freedom correction to remove the shrinkage bias of regularized estimators and conduct inference. However, this method requires that the observed samples are i.i.d., the covariates follow a mean zero Gaussian distribution, and reliable covariance matrix estimates for observed features are available. This approach struggles when (i) covariates are non-Gaussian with heavy tails or asymmetric distributions, (ii) rows of the design exhibit heterogeneity or dependencies, and (iii) reliable feature covariance estimates are lacking. To address these, we develop a new strategy where the debiasing correction is a rescaled gradient descent step (suitably initialized) with step size determined by the spectrum of the sample covariance matrix. Unlike prior work, we assume that eigenvectors of this matrix are uniform draws from the orthogonal group. We show this assumption remains valid in diverse situations where traditional debiasing fails, including designs with complex row-column dependencies, heavy tails, asymmetric properties, and latent low-rank structures. We establish asymptotic normality of our proposed estimator (centered and scaled) under various convergence notions. Moreover, we develop a consistent estimator for its asymptotic variance. Lastly, we introduce a debiased Principal Components Regression (PCR) technique using our SpectrumAware approach. In varied simulations and real data experiments, we observe that our method outperforms degrees-of-freedom debiasing by a margin.

Heckman Selection - Contaminated Normal Model

Heeju Lim, University of Connecticut

Co-authors: José Alejandro Ordoñez; Antonio Punzo; Victor Hugo Lachos Davila

Abstract: The Heckman selection model is one of the most popular econometric models in the analysis of data with sample selection. This model is designed to rectify sample selection biases based on the assumption of bivariate normal error terms. However, real data diverge from this assumption in the presence of heavy tails and/or mildly atypical observations. Recently, this assumption has been relaxed to more flexible models based on the Student's t-distribution, which has appealing statistical properties. This paper introduces a novel Heckman selection model using a bivariate contaminated normal distribution for the error terms. We present an efficient ECM algorithm for parameter estimation with closed-form expressions at the E-step based on truncated multinormal distribution formulas. The identifiability of the proposed model is also discussed, and its properties have been examined. Through simulation studies, we compare our proposed model with the normal and Student's t counterparts and investigate the finite-sample properties and the missing variation. Results obtained from two real data analyses showcase the usefulness and effectiveness of our model. The proposed algorithms are implemented in the R package HeckmanEM.

A Distribution-Free Mixed Integer Optimization Approach to Hierarchical Modelling of Clustered and Longitudinal Data

Madhav Sankaranarayanan, Harvard T.H. Chan School of Public Health

Co-authors: Intekhab Hossain, Tom Chen

Abstract: Recent advancements in Mixed Integer Optimization (MIO) algorithms, paired with hardware enhancements, have led to significant speedups in resolving MIO problems. These strategies have been utilized for optimal subset selection, specifically for choosing k features out of p in linear regression given n observations. In this paper, we broaden this method to facilitate cluster-aware regression, where selection aims to choose λ out of K clusters in a linear mixed effects (LMM) model with n_k observations for each cluster. Through comprehensive testing on a multitude of synthetic and real datasets, we exhibit that our method efficiently solves problems within minutes. Through numerical experiments, we also show that the MIO approach outperforms both Gaussian- and Laplace-distributed LMMs in terms of generating sparse solutions with high predictive power. Traditional LMMs typically assume that clustering effects are independent of individual features. However, we introduce an innovative algorithm that evaluates cluster effects for new data points, thereby increasing the robustness and precision of this model. The inferential and predictive efficacy of this approach is further illustrated through its application in student scoring and protein expression.

Robust and Ecient Integration of Blockwise Missing and Semi-Supervised Data

Yiming Li, Columbia University, NY

Co-authors: Xuehan Yang, Molei, Liu and Ying, Wei

Abstract: Data fusion is an important way to realize powerful and generalizable regression analyses with multiple sources. However, different capability of data collection across the sources has become a prominent issue in practice, to result in the block-wise missingness (BM) of covariates making the integrative regression challenging. Meanwhile, the high cost of obtaining gold-standard labels can cause the missingness of response on a large proportion of the samples, known as the semi-supervised (SS) problem. In this paper, we consider a new scenario in data fusion confronting both the BM and SS issues, and propose a novel Data-adaptive projecting Estimation approach for data FUsion in the SEmi-supervised setting (DEFUSE). Starting with a complete-data-only estimator, it involves two successive projection steps to reduce its variance without incurring any bias. Compared to existing approaches, DEFUSE is shown to achieve a two-fold improvement. First, it uses the BM labeled sample in a more efficient way resulting in better variance reduction. Second, it further incorporates the large unlabeled sample to enhance the estimation efficiency through score-projection. In the classic complete data scenario with correct outcome models, the unlabeled sample is known to be nuisance and cannot help to improve efficiency. Interestingly, in contrast to this, our work reveals a more essential role of the large unlabeled sample in the BM setting. These advantages of DEFUSE are justified in asymptotic and simulation studies. To illustrate its utility, we also apply our method for the risk modeling and inference of heart diseases with the MIMIC-III electronic medical record (EMR) data set.

Are Arbitration Eligible MLB Players Paid

Fairly?

Jon Gordon, Bentley University

Abstract: Major league baseball players' salaries are initially determined by the contract they sign after being drafted. They are not free to negotiate a new salary until they reach three years of major league service time. After the third year, they can negotiate their next contract with their current team. If they cannot reach an agreement with their club, their salary will be determined by an independent arbiter. These players are identified as "arbitration eligible". In this paper we will use both clustering techniques (such as DBSCAN, K-Means, TSNE) and regression techniques (such as KNN, Bagged Tree, and Random Forest) to predict the salary the arbitration players might expect to receive, using publicly available 2023 player statistics. The predicted salaries will then be compared to the actual amounts negotiated or awarded by the independent arbitration panel. The results show that certain categories of players are underpaid, but not all categories. The inability of the arbitration players to negotiate with any club does not appear to be as big of a detriment as expected.

Parallel Session 4 | 08:45 AM - 10:25 AM, May 23

Design and Analysis of Network-Based Studies to Inform Public Health and Education Policy (IS-11)

Chair: Tianyu Sun

Proposer: Ke Zhang, Department of Computer Science and Statistics, University of Rhode Island

Room: McHugh 201

Presenters: Ke Zhang; Ashley Buchanan; Samrachana Adhikari; Fei Fang

Power and Sample Size Calculations for Evaluating Spillover Effects in Networks with Non-Randomized Interventions

Ke Zhang, University of Rhode Island

Abstract: Network-based analyses are crucial for understanding social and epidemiological phenomena. Spillover effects, the effects of one's exposure on others' outcomes, are often of particular interest as they may lead to improvements in effectiveness of the interventions. Recent work developed methods for assessing spillover in network studies, but their statistical power implications remain largely unknown. In this work, we conducted a simulation study to investigate the impact of different design parameters on the statistical power for assessing spillover effects. The parameters include number of components, number of nodes, node degree, transitivity (i.e., global clustering coefficient), and effect size (i.e., true spillover effect). The results suggested that (1) power increased with more nodes or larger effect size, but did not necessarily change with more components when the number of nodes was fixed; (2) power decreased with a higher node degree or transitivity; and (3) power decreased significantly in a network with a dominant giant component (i.e., a group of nodes that contains a substantial portion of the network) compared to a more evenly distributed network. These findings can help in designing network-based studies of spillover to ensure the study has adequate power.

Assessing Diffusion and Contamination Effects of A Network-Randomized Intervention for HIV Prevention among People Who Inject Drugs

Ashley Buchanan, University of Rhode Island

Abstract: We used a potential outcomes (po) framework to assess contamination/diffusion measured by intervention knowledge. The control is inadvertently exposed to an intervention via contamination. Diffusion is purposeful exposure spread. HIV Prevention Trials Network 037 was a network-randomized trial among people who inject drugs. Indexes were recruited then their sexual/drug use partners. Intervention indexes were trained to diffuse behaviors/information to networks. A contamination effect compares average po of control participants exposed via contamination (vs not) for indexes and nonindexes. A diffusion effect is similarly defined for intervention nonindexes. The outcome was injection risk behaviors at follow-up. Risk ratios were estimated using a log-binomial model fit with GEEs adjusting for confounding. The intervention effect among nonindexes was 0.52 (95% confidence interval: 0.29, 0.92). Diffusion for intervention nonindexes was 0.71 (0.21, 2.38). Contamination for control indexes and nonindexes were 0.38 (0.13, 1.17) and 0.89 (0.41, 1.93). The intervention via contamination may have a weaker effect. Our method may help assess contamination effects in network-randomized studies.

Flexible Bayesian Framework to Estimate Causal Peer Influence Accounting for Latent Network Homophily

Samrachana Adhikari, NYU School of Medicine

Abstract: Despite the widespread interest in peer influence, identifying and accurately estimating causal peer influence effect using observational network data is often challenging due to confounding by multiple measured and unmeasured factors that lead to homophily. In many social networks it is not possible to observe or measure all homophilous attributes that lead to relational ties, resulting in latent homophily. Further, when there is confounding due to latent homophily, peer influence effect is not causally identified without additional assumptions. We propose an approach to estimate causal peer influence effects while adjusting for confounding due to latent homophily. Specifically, we present identification assumptions for the causal peer influence effect in the presence of latent homophily, extending previous work of McFowland III and Shalizi, and develop a flexible semi-parametric Bayesian hierarchical framework for estimation.

Average and Conditional Inward and Outward Spillovers of One Unit's Treatment under Net-

work Interference

Fei Fang, Yale University

Abstract: In a connected social network, users may have varying levels of influence on others when they themselves receive interventions. For example, giving an advertisement to a more influential person can have on average a greater impact on others' purchase decisions. Understanding and evaluating these effects can provide valuable insights for various applications such as targeting strategies in marketing and behavioral interventions in public health. Under a partial interference assumption, we define influence effects in two ways: i) the inward average spillover effect on a unit's outcome of a neighbor's treatment, and ii) the outward average spillover of a unit's treatment on their neighbors' outcomes. We investigate the comparison between the two causal effects in directed networks with different properties, including the conditions under which they are equivalent. Additionally, we develop Horvitz-Thompson estimators for assessing both effects, on average and conditioning on categorical covariates, as well as weighted least square estimators for these effects conditioning on continuous covariates. We derive design-based variance estimators and establish the consistency and asymptotic normality. Through simulations, we verify the empirical performance of our proposed estimators. Finally, we employ our approach to investigate inward and outward average and conditional spillover effects of an information session on the adoption of weather insurance among rice farmers in China.

Statistical Inference, Geometry and AI (IS-31)

Chair: Aritra Halder

Proposer: Aritra Halder and Didong Li, Drexel University, University of North Carolina

Room: McHugh 202

Presenters: Rajarshi Guhaniyogi; Didong Li; Hengrui Luo; Lu Zhang

Random Compression Matrices: From Distributed Learning to Manifold Regression

Rajarshi Guhaniyogi, Texas A & M University

Abstract: In this presentation, we will delve into the versatility of random compression matrices in facilitating accurate and efficient inference across various scientific domains. Our focus will be on their utilization in addressing two specific challenges: firstly, in the development of robust distributed Bayesian learning strategies for emulator data to predict water surge during hurricanes; and secondly, in Bayesian manifold regression for forecasting outdoor air pollution using remote sensing technology. In both cases, random compression matrices serve as crucial components, expediting Bayesian inference processes and enabling precise quantification of uncertainty.

The Impossible Triangle in Deep Generative Models: Complexity, Dimensionality, and Approximation

Didong Li, University of North Carolina at Chapel Hill

Abstract: Generative networks have shown remarkable success in learning complex data distributions, particularly in generating high-dimensional data from lower-dimensional inputs. While this capability is well-documented empirically, its theoretical underpinning remains unclear. One common theoretical explanation appeals to the widely-accepted manifold hypothesis, which suggests that many real-world datasets, such as images and signals, often possess intrinsic low-dimensional geometric structures. Under this manifold hypothesis, it is widely believed that to approximate a distribution on a d-dimensional Riemannian manifold, the latent dimension needs to be at least d or $d+1$. In this work, we show that this requirement on the latent dimension is not necessary by demonstrating that generative networks can approximate distributions on d-dimensional Riemannian manifolds from inputs of any arbitrary dimension, even lower than d, using the concept of space-filling curves. This approach, in turn, requires an increased complexity of the deep neural networks through expanded layers and neurons. Our findings thus challenge the conventional belief about the relationship between input dimensionality and the ability of generative networks to model data distributions. This novel insight not only corroborates the practical effectiveness of generative networks in handling complex data structures, but also underscores a critical trade-off between dimensionality and model complexity.

Sharded Bayesian Additive Regression Trees

Hengrui Luo, Rice University

Abstract: For non-smooth regression problems, we investigate the scalability issue with the Bayesian Additive Regression Tree (BART) model and discuss enhancing computational efficiency through data shards. We proposed the Sharded Bayesian Tree (SBT) model further advances scalability in Bayesian tree models, integrating innovative concepts of randomized sharding and intersection partitioning trees for efficient large-scale data handling. We introduce a randomization auxiliary variable and a sharding tree to decide partitioning of data, and fit each partition component to a sub-model using Bayesian Additive Regression Tree (BART). By observing that the optimal design of a sharding tree can determine optimal sharding for sub-models over a product space, we introduce an intersection tree structure to completely specify both the sharding and modeling using only tree structures. In addition to experiments, we also derive the theoretical optimal weights for minimizing posterior contractions and prove the worst-case complexity of SBT. These innovative modeling techniques offer a scalable solution in big-data regression analysis and opens the door to using sharding in modern nonparametric and machine learning models.

Bayesian Geostatistics Using Predictive Stacking

Lu Zhang, University of Southern California

Abstract: In this talk, we present Bayesian predictive stacking for geostatistical models, where the primary inferential objective is to provide inference on the latent spatial random field and conduct spatial predictions at arbitrary locations. We exploit analytically tractable posterior distributions for regression coefficients of predictors and the realizations of the spatial process conditional upon process parameters. We subsequently combine such inference by stacking these models across the range of values of the hyper-parameters. We devise stacking of means and posterior densities in a manner that is computationally efficient without resorting to iterative algorithms such as Markov chain Monte Carlo (MCMC) and can exploit the benefits of parallel computations. We offer novel theoretical insights into the resulting inference within an infill asymptotic paradigm and through empirical results showing that stacked inference is comparable to full sampling-based Bayesian inference at a significantly lower computational cost.

Modern Methods for Analysis of High-Dimensional Data in Biomedical Research (IS-33)

Chair: Wenrui Li

Proposer: Wenrui Li, University of Pennsylvania

Room: McHugh 205

Presenters: Xin Ma; Jun Young Park; Hai Shu; Wenrui Li

High-Dimensional Measurement Error Models for Lipschitz Loss

Xin Ma, Columbia University

Abstract: Recently emerging large-scale biomedical data pose exciting opportunities for scientific discoveries. However, the ultrahigh dimensionality and non-negligible measurement errors in the data create difficulties in estimation. There are limited methods for high-dimensional covariates with measurement errors, that usually require moments of the noise distribution to fit the working model and are restricted to generalized linear models (GLM). In this work, we develop measurement error models involving high-dimensional covariates with correlated sub-Gaussian measurement errors for a class of Lipschitz loss functions that go beyond GLM family, and encompass logistic regression, hinge loss and quantile regression. Our estimator is designed to minimize the L1 norm among all estimators in suitable feasible sets, without requiring any knowledge of the noise distribution. Subsequently, we generalize these estimators to a Lasso analog version that is computationally scalable to higher dimensions. We derive theoretical guarantees of finite sample statistical error bounds and sign consistency, even when the dimensionality increases exponentially with the sample size. Extensive simulation studies demonstrate superior performance compared to existing methods in classification and quantile regression problems. Applications to two classification examples with real imaging data illustrate improved classification accuracy under proposed approaches, and reliability in identifying brain features related to clinical phenotypes.

Promises of Covariance Modeling in High-Dimensional Neuroimaging Data

Jun Young Park, University of Toronto

Abstract: Neuroimaging studies arouse interest in statistics for their fruitfulness in data types and re-

search questions, but recent studies (e.g., Marek et al. (2022) Nature) showed that the signal-to-noise level ratio and replication rate are poor in neuroimaging. This necessarily requires a careful look at underlying multivariate patterns to gain efficiency in modeling with limited sample sizes and high-dimensional features. This talk outlines how brain-wise covariance modelling would help improve statistical power in such scenarios, with a focus on spatial-extent inference. Our method, CLEAN, proposes a unified methodology to address various research in neuroimaging, including testing and localizing (i) intermodal correlations, (ii) test-retest reliability, and (iii) heritability, while keeping it computationally efficient for its practical utility. We use data-driven simulations from the Human Connectome Project (HCP) and the Philadelphia Neurodevelopment Cohort (PNC) study to show its promising empirical performance.

Deepfdr: A Deep Learning-Based False Discovery Rate Control Method for Neuroimaging Data

Hai Shu, New York University

Abstract: Voxel-based multiple testing is widely used in neuroimaging data analysis. Traditional false discovery rate (FDR) control methods often ignore the spatial dependence among the voxel-based tests and thus suffer from substantial loss of testing power. While recent spatial FDR control methods have emerged, their validity and optimality remain questionable when handling the complex spatial dependencies of the brain. Concurrently, deep learning methods have revolutionized image segmentation, a task closely related to voxel-based multiple testing. In this paper, we propose DeepFDR, a novel spatial FDR control method that leverages unsupervised deep learning-based image segmentation to address the voxel-based multiple testing problem. Numerical studies, including comprehensive simulations and Alzheimer's disease FDG-PET image analysis, demonstrate DeepFDR's superiority over existing methods. DeepFDR not only excels in FDR control and effectively diminishes the false nondiscovery rate, but also boasts exceptional computational efficiency highly suited for tackling large-scale neuroimaging data.

Graph-Guided Bayesian Factor Model for Integrative Analysis of Multi-Modal Data with Noisy Network Information

Wenrui Li, University of Pennsylvania

Abstract: There is a growing body of literature on factor analysis that can capture individual and shared structures in multi-modal data. However, few of these approaches incorporate biological knowledge such as functional genomics and functional metabolomics. Graph-guided statistical learning methods that can incorporate knowledge of underlying networks have been shown to improve predication and classification accuracy, and yield more interpretable results. Moreover, these methods typically use graphs extracted from existing databases or rely on subject matter expertise which are known to be incomplete and may contain false edges. To address this gap, we propose a graph-guided Bayesian factor model that can account for network noise and identify globally shared, partially shared and modality-specific latent factors in multi-modal data. Specifically, we use two sources of network information, including the noisy graph extracted from existing databases and the estimated graph from observed features in the dataset at hand, to inform the model for the true underlying network via a latent scale modeling framework. This model is coupled with the Bayesian factor analysis model with shrinkage priors to encourage feature-wise and modal-wise sparsity, thereby allowing feature selection and identification of factors of each type. We develop an efficient Markov chain Monte Carlo algorithm for posterior sampling. We demonstrate the advantages of our method over existing methods in simulations, and through analyses of gene expression and metabolomics datasets for Alzheimer's disease.

Bridging The Knowledge Gap: Advances in Pediatric Extrapolation (IS-46)

Chair: Vickie (Yuanye) Zhang;

Proposer: Vickie (Yuanye) Zhang, Servier BioInnovation

Room: McHugh 206

Presenters: James Rogers; Vickie (yuanye) Zhang; Yanyan Zhu

Formalizing "Similarity" in Pediatric Extrapolation Plans Using Causal Selection Graphs

James A. Rogers, Metrum Research Group

Abstract: In the regulatory context for pediatric extrapolation considerations regarding "similarity of disease and response to treatment between reference and target pediatric population" constitute the basis for "extrapolations concepts", and subsequently for "extrapolation plans" (EMA 2022). In this context, we propose that "similarity" be mathematically formalized as statistical exchangeability. More specifically, we propose that the task of accurately translating between informal notions of similarity and formal notions of exchangeability is facilitated by the use of "selection graphs", an extension of causal directed acyclic graphs (DAGs) developed by Pearl and Bareinboim (2014) to elucidate the logic of extrapolation. As such, our proposal may be understood as a consulting tool (and not as a statistical methodology per se) that a statistician or pharmacometrician may use to verify that a given extrapolation plan implements a version of exchangeability that accurately reflects the baseline assumptions of key (quantitative and non-quantitative) stakeholders. We illustrate the use of this "consulting tool" in the context of published examples of pediatric extrapolation. European Medicines Agency. ICH guideline E11A on pediatric extrapolation, Step 2b. 2022. Judea Pearl, Elias Bareinboim "External Validity: From Do-Calculus to Transportability Across Populations," Statistical Science, Statist. Sci. 2014 Nov; 29(4), 579-595.

Recent Use of Pediatric Extrapolation in Pediatric Drug Development in US

Vickie (yuanye) Zhang, Servier BioInnovation

Abstract: The regulatory standards of the United States Food and Drug Administration (FDA) require substantial evidence of effectiveness from adequate and well-controlled trials, for drugs developed in both adults and children. However, when it is not feasible or ethical to conduct such trials in children, relying on extrapolation may be acceptable. Historically, FDA's extrapolation approach was mainly based on draft guidance published in 2014 as category of full, partial and no extrapolation. European Medicines Agency (EMA) took a different view on pediatric extrapolation. To better understand the use of extrapolation to support pediatric drug development and approval, we reviewed the pediatric labeling changes published by FDA, focusing on the labeling updates between 1/1/2015 and 7/31/2021, the period where the extrapolation approach is in transition to harmonize with EMA. Within this time window, among the 265 drugs and biological products with pediatric labeling changes, 169 (63.8%) were identified where extrapolation was used. This includes 64 (24.2%) labeling changes, where full extrapolation was used and 105 (39.6%) labeling changes, where partial extrapolation was used. The major disease areas that extrapolation was used include neuroscience (40/53, 75.5%), and infectious disease (20/28, 71.4%). The change of extrapolation approaches was identified in terms of source population beyond the use of adult as well as allowing extrapolation from clinical trials conducted in the same drug class. The use of extrapolation increased the rates of new and expanded pediatric indication in the period. This review gives the most recent landscape of pediatric labeling changes using extrapolation. With the released ICH E11A guidance in April 2022, the paper also provides insights for future pediatric drug development programs.

Mechanism of Action (Moa)-Based Extrapolation: Bayesian Re-Design of A New Pediatric Trial via Borrowing Information From The Concurrent Adult Trials and Historical Pediatric and Adult Trials From The Same Class of Drugs

Yanyan Zhu, University of Connecticut

Abstract: Pediatric trials pose unique and challenging circumstances for several reasons: a small patient population, limited physiological data, and ethical complexity. Consequently, pediatric drug development often lags behind that of adults post-drug approvals. To address this issue, ICH E11A proposes a complementary strategy, known as pediatric extrapolation. The approach involves assessing the relevance of existing information from the adult/reference population to the target/pediatric population. It focuses on aspects of similarity in the disease, drug pharmacology and clinical response to treatment. The aim is to identify the gaps or level of uncertainty that must be addressed to extend conclusions regarding adequate evidence of efficacy and safety. Despite the inherent challenges, extrapolation based on Mechanism-of-action (MOA) emerges as a viable option in this context, e.g., extrapolation from adult to the pediatric population using clinical trials within the same drug class can be leveraged. In this paper, we propose a new randomized Bayesian test for designing a single-arm superiority trial within a general Bayesian decision rule-based framework. The analytic form of the test statistic for binary primary outcomes is derived and a search algorithm to achieve the exact type I error is developed. The desirable theoretical properties of the proposed test are established. Several priors are investigated for leveraging multiple historical data.

The type I error and the power are computed exactly without resorting Monte Carlo sampling. Furthermore, an analytical procedure is devised to determine the amount of borrowing from the historical data in order to control a pre-specified inflation level of Type I error. The usefulness of the proposed methodology is further demonstrated via re-designing a new pediatric superiority trial borrowing information data from concurrent adult trials and historical pediatric and adult trials from the same class of drugs.

Opportunities and Challenges in The Use of Electronic Health Records for Making Informed Clinical Decisions (IS-50)

Chair: TBD

Proposer: Shuangge (Steven) Ma, Lei Yan, Yale School of Public Health

Room: McHugh 301

Presenters: Lei Yan; Kelson Zawack; Anita Wang; Yuan Huang

Identifying Veterans Who Benefit From Nirmatrelvir-Ritonavir: Opportunities and Challenges in The Use of Electronic Health Records

Lei Yan, Yale School of Public Health

Abstract: Nirmatrelvir-ritonavir is recommended to treat nonhospitalized persons with mild-to-moderate COVID-19. Although randomized controlled trials demonstrated the efficacy of nirmatrelvir-ritonavir in reducing COVID-19-related hospitalization or death, the absolute benefit in subsequent clinical and observational studies has been smaller due to uptake of vaccination and circulation of different viral variants. We conducted a target trial emulation study using electronic health records (EHRs) from the Veterans Heath Administration to identify which eligible persons meaningfully benefit from antiviral treatment. With over 100,000 outpatient veterans diagnosed with COVID-19, we compared the 30-day risk of death or hospitalization between those who received nirmatrelvir-ritonavir and those who didn't for the entire cohort, as well as among patient subgroups. We found that nirmatrelvir-ritonavir was associated with a reduced 30-day risk in persons aged 65 years or older, those at the highest risk predicted by an ensemble learning model, and immunocompromised individuals. The richness of EHR data allows us to define subgroups of interest more accurately, which is clinically important but too costly for clinical trials. The challenges of using EHR data, including integrating external data, will also be discussed.

Examining Racial and Ethnic Disparities in Continuous Glucose Monitor Prescriptions Using Electronic Health Records

Kelson Zawack, Department of Biomedical Informatics and Data Science, Yale School of Public Health, New Haven, CT Veterans Affairs Cooperative Studies Program Clinical Epidemiology Research Center (CSP CERC), Veterans Affairs Connecticut Healthcare System, West Haven, CT

Abstract: Continuous glucose monitors (CGM) provide real time data on blood glucose levels and have been shown to help patients better manage their diabetes. As a result, professional society guidelines recommend the use of CGM by patients with diabetes on insulin therapy. Previous work, however, has identified racial and ethnic disparities in health outcomes generally and the prescription of CGM specifically. This study endeavors to determine whether such disparities in CGM prescribing exist in the Veterans Health Administration system using electronic health record (EHR) data. Because of the high cost of insulin in the private market and the fact that insulin is covered for veterans filling their prescriptions for insulin at the VA, veterans commonly seek their diabetes care from the VA. Furthermore, the large number of covariates captured in electronic health records make them an essential source of information for untangling outcomes with complex etiologies like health disparities. EHR data, however, also introduces complexities into the analysis as not all clinically relevant variables are perfectly captured and working proxies must be developed. After accounting for age, sex, service connection status, geographic region, rurality, area deprivation index, diabetes type, history of severe hypoglycemia, type of insulin therapy, hemoglobin A1c, endocrinologist visit count, and the rate of cgm prescribing at a veteran's home VA station this study finds that Black and Hispanic veterans receive CGM at lower rates than white veterans. This evidence continues to build the case for interventions that produce more equitable care for all patients.

Enhancing Patient Representation Learning From Electronic Health Records through Predicted Family Relations

Zuoheng Wang, Yale University

Abstract: Artificial intelligence and machine learning are powerful tools for analyzing electronic health records (EHRs) in healthcare research. Despite the recognized importance of family health history, patients are often treated as independent samples in traditional analyses, overlooking family relations. To address this gap, we present ALIGATEHR, which models predicted family relations in a graph attention network integrated with a medical ontology. Taking disease risk prediction as a use case, we first demonstrate that explicitly modeling family relations significantly improves predictions across the disease spectrum. We then show how ALIGATEHR's attention mechanism successfully captures genetic aspects of diseases using only EHR diagnosis data. Finally, we use ALIGATHER to successfully distinguish the two main inflammatory bowel disease subtypes (Crohn's disease and ulcerative colitis). Our results highlight that family relations should not be overlooked in EHR research and illustrate ALIGATEHR's great potential for improving patient representation learning for predictive and descriptive modeling of EHRs.

Leveraging Electronic Health Records to Make Informed Decisions Based on Real-Word Data

Yuan Huang, Yale University

Abstract: These three talks will provide an overview of approaches and methods that can be used to extract real-word data from electronic health records (EHRs). Based on the data presented, we will have a panel discussion to engage the audience to share their interpretation and ideas. In particular, we will ask the attendees to share their own experiences regarding the method development and analysis of large-scale and heterogeneous datasets such as EHRs. This will allow us to further expand the engagement of the attendees working in different research areas. Indeed, we expect that attendees with different scientific backgrounds will be interested in the broad implications of the topics discussed in this session. Among them, investigators focused on Big Data Analytics are expected to have the most interest in joining this session to learn more about the challenges and opportunities available for real-word data research.

Beyond Independent and Identically Distributed: Models for Non-Standard Data (IS-57)

Chair: Maryclare Griffin

Proposer: Maryclare Griffin, University of Massachusetts Amherst

Room: McHugh 305

Presenters: Rebecca Kurtz-garcia; Carlos Soto; Nathan Wycoff; Qian Zhao

Bandwidth Selection for Zero Lugsail Kernels

Rebecca Kurtz-garcia, Smith College

Abstract: Test statistics, confidence intervals, and p-values all typically rely on an estimate for variance. For data sets that are not independent and identically distributed (iid) caution must be used when selecting a variance estimator. If the dependence structure is unknown but stationary, a robust long run variance (LRV) estimator can be used which can handle a wide variety of scenarios. Spectral variance (SV) estimators are one of the most common LRV estimation methods, but they suffer from a negative bias in the presence of positive correlation. An alternative zero lugsail estimator has been proposed to combat this issue which has a zero asymptotic bias regardless of correlation. Both SV and zero lugsail estimators rely on a bandwidth parameter, a critical component for the estimation process. Currently no guidelines exist for selecting a bandwidth for the zero lugsail estimator. We propose an optimal bandwidth rule for zero lugsail estimators when relying on nonstandard limiting distributions. With this procedure we can greatly improve bias, account for variability, and obtain an estimator optimized for inference.

Representation of Chromosome Conformations Using A Shape Alphabet Across Modeling Methods

Carlos Soto, University of Massachusetts Amherst

Abstract: Despite enormous structural variability exhibited in 3D chromosomal conformations at a global scale, there is a significant commonality of structures visible at smaller, local levels. We hypothesize that chromosomal conformations are representable as concatenations of a handful of prototypical shapelets, termed shape letters. This is akin to expressing complicated sentences in a language using only a small set

of letters. Our goal is to organize the vast variability of 3D chromosomal conformation by constructing a set of predominant shape letters, termed a shape alphabet, using statistical shape analysis of curvelets taken from training conformations. This paper utilizes conformations generated from Integrative Genome Modeling to develop a shape alphabet as follows: it first segments 3D conformations into curvelets according to their Topologically Associated Domains. It then clusters these segments, estimates mean shapes, and refines and reorders these shapes into a Chromosome Shape Alphabet. The paper demonstrates effectiveness of this construction by successfully representing independent test conformations taken from IGM and other methods such as SIMBA3D, both symbolically and structurally, using the constructed alphabet.

Proximal Iteration for Non-Gaussian Adaptive Lasso

Nathan Wycoff, Georgetown University

Abstract: Regression with an ℓ_1 penalty, or Lasso regression, is a major tool in the practice of statistics due to its capability to perform automatic variable selection by thresholding model parameters to zero. Furthermore, inference is possible on an arbitrary smooth cost functions augmented by an ℓ_1 penalty by way of proximal gradient methods, which efficiently deal with the absolute value's nonsmoothness. However, one drawback of the ℓ_1 penalty is bias: nonzero parameters are underestimated in absolute value, motivating techniques such as the Adaptive Lasso which allow each parameter its own penalty coefficient. A major question left open is then how to choose these penalty coefficients, particularly for complex models. In this talk, we'll develop a proximal gradient method which can jointly optimize both the parameters and penalty coefficients, treating the latter as additional decision variables to be learned in a Maximum a Posteriori manner. In addition to reducing bias in estimates, this procedure also allows us to encourage arbitrary structure in the sparsity by imposing an appropriate prior on the penalty coefficients. We compare our method to software implementing specific sparisty structures on synthetic and real datasets with generalized linear models, where we find our method to be competitive in terms of both speed and accuracy while being far more general. We then consider nonlinear models consisting of two social science case studies: first a deep active subspace classifier of COVID-19 vaccination behavior and second to a model of international migration which shows the applicability of our method to general non-Gaussian likelihoods.

Beta-Trees — Multivariate Histograms with Confidence Statements

Qian Zhao, University of Massachusetts, Amherst

Abstract: Multivariate histograms are difficult to construct due to the curse of dimensionality. Motivated by k-d trees in computer science, we show how to construct an efficient data-adaptive partition of Euclidean space that possesses the following two properties: (1) with high confidence the distribution from which the data are generated is close to uniform on each rectangle of the partition; and (2) finite sample simultaneous confidence intervals can be provided for the probabilities of each rectangle in the partition. The method produces confidence intervals whose widths depend only on the probability content of the rectangles and not on the dimensionality of the space, thus avoiding the curse of dimensionality. Moreover, the widths essentially match the optimal widths in the univariate setting. The simultaneous validity of the confidence intervals allows us to use this construction, which we call Beta-trees, for various data-analytic purposes. We illustrate this by using Beta-trees for visualizing and for multivariate mode-hunting of the flow cytometry data.

Innovations in Statistical Machine Learning: Methodology and Inference Theories (IS-63)

Chair: Minghui Chen

Proposer: Minge Xie, Rutgers University

Room: McHugh 101

Presenters: Colin Wu; Zhenyu Wang; Linjun Zhang

Knowledge-Guided Statistical Machine Learning for Longitudinal Biomedical Studies

Colin O. Wu, Office of Biostatistics Research, National Heart, Lung and Blood Institute, National Institutes of Health

Abstract: In large epidemiological studies, a comprehensive analysis of a disease process often includes several statistical sub-models with time-varying risk factors (i.e. functional covariates) and time-to-event disease outcomes. A useful prediction model for the

disease should be constructed by incorporating all the influential covariates and functional features (i.e. "history") of the longitudinal risk factors, so that meaningful clinical interpretations for the disease prediction model could be obtained. Existing statistical machine learning methods lack a systematic approach for incorporating all the influential functional covariates and sub-models in a clinically meaningful way. We develop a "knowledge-guided statistical machine learning" (KGSML) procedure to construct a statistical model for predicting the time-to-event and/or disease outcomes with time-varying risk factors. This KGSML procedure sequentially combines flexible statistical machine learning methods, such as nonparametric regression, spline-based subject-specific best linear unbiased prediction (BLUP), random survival forest, survival regression models and variable selection. Application of this KGSML procedure to a landmark epidemiological study of NHLBI, the Coronary Artery Risk Development in Young Adults (CARDIA) Study, demonstrates that early treatment of cardiovascular risk factors during young adulthood could dramatically reduce the risk of incident cardiovascular disease (CVD) in midlife. This finding leads to novel insights into the potential benefit of early primary prevention strategies for reducing the global CVD burden. We finally discuss some future research directions of statistical machine learning for longitudinal studies in biomedical research.

Principled Random Forests: Uncertainty Quantification for Tree Structure Models with Robustness to Stratification Errors

Zhenyu Wang, Rutgers University

Abstract: Tree models are one of the machine learning models used to estimate the conditional mean E[Y |X], known for their interpretability and straightforward application. However, a significant limitation of these models is the lack of valid inferential tools, especially in finite sample scenarios. This work introduces Principled Random Forests (PRF), a novel methodology that uses synthetic random errors to help address inference problems. Our theoretical development shows that the PRF method has inference performance guarantees under mild conditions. The PRF method is further generalized to and particularly adept at analyzing conditional average treatment effects within the realm of causal inference, accommodating heterogeneity across different subgroups. A novel filtering technique is also proposed and the enhancement significantly improves the inference efficiency of PRF by reducing the length of confidence intervals. Both numerical simulations and real-world applications demo nstrate the effectiveness of our method, showcasing its potential to advance the utility of tree models in complex analytical tasks.

Finite-Sample and Distribution-Free Fair Classification: Optimal Excess Risk-Fairness Trade-Off and The Cost of Group-Blindness

Linjun Zhang, Rutgers University

Abstract: Algorithmic fairness in machine learning has attracted significant attention recently, yet the impact of group fairness on excess risk remains unclear. Despite the widespread adoption of group-blindness to promote fairness, it is effectiveness is uncertain. In this work, we explore the influence of fairness and group-blindness in the context of binary classification with group fairness constraints. Specifically, we propose a unified framework for fair classification with excess risk control and distribution-free and finite-sample fairness guarantees for various group fairness notions in both group-aware and group-blind scenarios. Moreover, for binary sensitive attributes, a minimax excess risk lower bound is provided, confirming the minimax optimality of the proposed algorithm. The minimax excess risk reveals the inherent trade-off between excess risk and fairness, and uncovers the inevitable cost of group-blindness, which may lead to constant excess risk in extreme cases. Through simulation studies and real data analysis, we illustrate the superior performance of our algorithm compared to existing methods and also provide empirical evidence for our theoretical findings.

Student Paper 4 (S4)

Chair: TBD

Organizer: Neil Spencer

Room: McHugh 306

Reliable Multivariate Deep Regression Using Moment-Matching Prior Networks

Qingyi Pan, Tsinghua University & Purdue University

Co-authors: Ruqi Zhang

Abstract: When deep neural networks are deployed in high-stakes applications, uncertainty estimation is crucial for reliable predictions and decision-making. Despite rich studies in univariate deep regression, multivariate deep regression with accurate uncertainty estimation, especially concerning the covariance matrix, remains largely unexplored. In this paper, we propose a scalable evidential prior to capturing both aleatoric and epistemic uncertainty, including the correlation of the multivariate response vector. Our method formulates a hierarchical probabilistic framework where the evidential prior is fitted using samples generated by a neural network based on moment-matching. Extensive empirical results on real-world multivariate regression tasks demonstrate that our method provides accurate prediction and uncertainty estimation with minimal computational overhead, significantly outperforming existing methods.

Minimax Optimal Goodness-Of-Fit Testing with Kernel Stein Discrepancy.

Omar Hagrass, The Pennsylvania State University

Co-authors: Bharath Sriperumbudur and Krishna Balasubramanian

Abstract: We explore the minimax optimality of goodness-of-fit tests on general domains using the kernelized Stein discrepancy (KSD). The KSD framework offers a flexible approach for goodness-of-fit testing, avoiding strong distributional assumptions, accommodating diverse data structures beyond Euclidean spaces, and relying only on partial knowledge of the reference distribution, while maintaining computational efficiency. We establish a general framework and an operator-theoretic representation of the KSD, encompassing many existing KSD tests in the literature, which vary depending on the domain. We reveal the characteristics and limitations of KSD and demonstrate its non-optimality under a certain alternative space, defined over general domains when considering χ^2-divergence as the separation metric. To address this issue of non-optimality, we propose a modified, minimax optimal test by incorporating a spectral regularizer, thereby overcoming the shortcomings of standard KSD tests. Our results are established under a weak moment condition on the Stein kernel, which relaxes the bounded kernel assumption required by prior work in the analysis of kernel-based hypothesis testing. Additionally, we introduce an adaptive test capable of achieving minimax optimality up to a logarithmic factor by adapting to unknown parameters. Through numerical experiments, we illustrate the superior performance of our proposed tests across various domains compared to their unregularized counterparts.

Mode-Wise Principal Subspace Pursuit and Matrix Spiked Covariance Model

Runshi Tang, University of Wisconsin-Madison

Co-authors: Ming Yuan, Anru Zhang

Abstract: This paper introduces a novel framework called Mode-wise Principal Subspace Pursuit (MOP-UP) to extract hidden variations in both the row and column dimensions for matrix data. To enhance the understanding of the framework, we introduce a class of matrix-variate spiked covariance models that serve as inspiration for the development of the MOP-UP algorithm. The MOP-UP algorithm consists of two steps: Average Subspace Capture (ASC) and Alternating Projection (AP). These steps are specifically designed to capture the row-wise and column-wise dimension-reduced subspaces which contain the most informative features of the data. ASC utilizes a novel average projection operator as initialization and achieves exact recovery in the noiseless setting. We analyze the convergence and non-asymptotic error bounds of MOP-UP, introducing a blockwise matrix eigenvalue perturbation bound that proves the desired bound, where classic perturbation bounds fail. The effectiveness and practical merits of the proposed framework are demonstrated through experiments on both simulated and real datasets. Lastly, we discuss generalizations of our approach to higher-order data.

Transfer Learning for Covariance Matrix Estimation: Optimality and Adaptivity

Dongwoo Kim, The Wharton School, University of Pennsylvania

Abstract: This study explores the transfer learning framework for the estimation of covariance matrices. We exploit a bandable structure to address the complexities of high-dimensional covariance matrices, and employ auxiliary observations that exhibit a similar yet distinct covariance matrix. Our contributions include introducing a novel estimation technique utilizing a block tridiagonal operator and establishing the minimax rate of convergence under the spectral norm. Our results unveil a phase transition phenomenon that underscores the effectiveness of transfer learning and the utilization of source samples. In terms of practical application, we present a data-driven algorithm that dynamically adapts to unknown model parameters. While this algorithm may not achieve the optimal rate of convergence, it represents the best

possible approach under the adaptivity constraints. Any attempt to accelerate the convergence rate of our algorithm incurs additional costs. Notably, the proposed algorithm is designed to minimize the cost of adaptation, supported by the concept of the adaptive rate of convergence. The theoretical insights are further substantiated by a simulation study, which corroborates the practicality and efficiency of our algorithm.

Optimizer's Information Criterion: Dissecting and Correcting Bias in Data-Driven Optimization

Tianyu Wang, Columbia University

Co-authors: Garud Iyengar, Henry Lam

Abstract: In data-driven optimization, the sample performance of the obtained decision typically incurs an optimistic bias against the true performance, a phenomenon commonly known as the Optimizer's Curse and intimately related to overfitting in machine learning. We develop a general approach that we call Optimizer's Information Criterion (OIC) to correct this bias. OIC generalizes the celebrated Akaike Information Criterion from the evaluation of model adequacy, used primarily for model selection, to objective performance in data-driven optimization which is used for decision selection. Our approach analytically approximates and cancels out the bias that comprises the interplay between model fitting and downstream optimization. As such, it saves the computation need to repeatedly solve optimization problems in cross-validation, while operates more generally than other bias approximating scheme. We apply OIC to a range of data-driven optimization formulations comprising empirical and parametric models, their regularized counterparts, and furthermore contextual optimization. Finally, we provide numerical validation on the superior performance of our approach under synthetic and real-world datasets.

Parallel Session 5 | 02:00 PM - 03:40 PM, May 23

Recent Advances in Network Analysis: Theory and Applications (IS-10)

Chair: Panpan Zhang

Proposer: Panpan Zhang, Vanderbilt University Medical Center

Room: McHugh 202

Presenters: Yuanjia Wang; Hosam Mahmoud; Laura Forastiere; Shiying Xiao

Identifying Temporal Pathways Using Biomarkers in The Presence of Latent Components

Yuanjia Wang, Columbia University

Abstract: Time series data collected from a network of random variables are useful for identifying temporal pathways among the network nodes. Observed measurements may contain multiple sources of signals and noises, including Gaussian signals of interest and non-Gaussian noises, including artifacts, structured noise, and other unobserved factors (e.g., genetic risk factors, disease susceptibility). Existing methods, including vector autoregression (VAR) and dynamic causal modeling do not account for unobserved non-Gaussian components. Furthermore, existing methods cannot effectively distinguish contemporaneous relationships from temporal relations. In this work, we propose a novel method to identify latent temporal pathways using time series biomarker data collected from multiple subjects. The model adjusts for the non-Gaussian components and separates the temporal network from the contemporaneous network. Specifically, an independent component analysis (ICA) is used to extract the unobserved non-Gaussian components, and residuals are used to estimate the contemporaneous and temporal networks among the node variables based on the method of moments. The algorithm is fast and can easily scale up. We derive the identifiability and the asymptotic properties of the temporal and contemporaneous networks. We demonstrate superior performance of our method by extensive simulations and an application to a study of attention-deficit/hyperactivity disorder (ADHD), where we analyze the temporal relationships between brain regional biomarkers.

Bernoulli Convolution of The Depth of Nodes in Recursive Trees with Generalized Affinities

Hosam Mahmoud, The George Washington University

Abstract: It has been reported in some sources that the depth of insertion of nodes in recursive trees grown with certain node affinities can be represented as a sum of independent Bernoulli random variables. In this presentation, we demonstrate that the presence of a Bernoulli convolution covers a much broader range of affinities beyond the uniform affinity, and the power weights. Further, we pursue sufficient conditions to have associated normal regimes and compare them to approximations via Poisson distributions. We give several illustrative examples, where we look at the rates of convergence, too.

Functional Connectivity Analysis in Brain Networks: A Statistical Review and Application to Alzheimer's Disease Data

Shiying Xiao, University of Connecticut

Abstract: Network-based functional connectivity analysis has emerged as a powerful tool for exploring interactions among brain regions. In recent years, brain network models based on precision matrix estimation have garnered considerable attention for analyzing functional connectivity in brain imaging data. Nevertheless, a thorough investigation into their practical effectiveness has been lacking. In this paper, we present a comprehensive statistical review of precision matrix estimation methods and their application in brain network modeling with Alzheimer's disease (AD) data. With the theoretical foundations of Gaussian graphical models and their relevance to functional connectivity analysis as a backdrop, a fine-grained review of various estimation methods is presented, including the graphical lasso (glasso), glasso with ridge penalty, graphical elastic net, adaptive glasso, SCAD, MCP, CLIME and TIGER. Following the review, we showcase the practical application of these methods in brain network modeling based on AD data. Using the AD data from the Tennessee Alzheimer's Project (TAP), we illustrate and compare their utility in identifying alterations in brain networks associated with AD pathology.

Statistics and AI In Finance: New Opportunities and Challenges (IS-12)

Chair: Yang Liu

Proposer: Yang Liu, Upstart Network, Inc.

Room: McHugh 101

Presenters: Filipp Shelobolin; Leander Eberhard; Yang Liu; Chaoyu Yuan; Claudio Antonini

Explaining Complex Machine Learning Models Using Surrogate Methodologies

Filipp Shelobolin, Upstart Network, Inc.

Abstract: The increasing effect of machine learning (ML) decisions on our lives requires more transparency of their inner workings. Accurate explanations of high-impact ML decisions such as denial of loan or job applications can provide valuable recourse to affected stakeholders and help model builders identify undesirable biases. However, much of the recent ML boom is due to growing complexity of models, which in turn makes them more difficult to explain. In the case that the features of a complex model are interpretable, a surrogate model methodology coupled with local feature attribution methods can be used to explain a model prediction in a model-agnostic way. In this talk, we will give a brief literature review of existing feature attribution methods and explain the setup of a surrogate model methodology. Then, we will explain the many limitations, challenges, and future opportunities of such approaches.

Algorithmic Fairness in Lending

Leander Eberhard, Upstart Network Inc.

Abstract: A key component in AI lending is ensuring the decisioning models treat applicants from all protected demographic groups fairly. In the tightly regulated financial industry, it is crucial to identify and avoid biases that could arise from new technologies. This discussion will explore various metrics used in the industry to asses the fairness of lending decisions, as well as methods to mitigate any detected disparities. We will focus in particular on the pros and cons of different fairness metrics, challenges in obtaining and using demographic information, and specific methodologies to improve the fairness of ML models.

Building An Effective Direct Mail Marketing Channel in Lending Leveraging Machine Learning

Yang Liu, Upstart Network, Inc.

Abstract: Marketing by physical mails may sound a bit outdated in this digital era, yet it is still proven to be highly effective and widely used in various industries. Direct mail marketing is often used to reach prospective customers in large scale, increase the brand awareness, and boost product sales. Typically, a direct mail model often built to target the people with highest response rates for a direct mail campaign, while this may not lead to the highest return of investment (ROI). With the development of artifical intelligence (AI) and its uses in lending, both new opportunities and challenges arise for building an effective marketing channel with direct mail. In this talk, we will discuss some of the processes and practical considerations of building a modern direct mail marketing channel leveraging machine learning techniques in a consumer finance context.

Event History Analysis in Lending

Chaoyu Yuan, Upstart Network, Inc.

Abstract: One of primary objective in the application of lending is to ascertain the relationship between different interventions and borrowers' response. Some traditional regression methods may ignore the hetergeneity among borrowers due to the lack of latent factors that have impact on lending outcomes. This discussion will explore some novel statistical techniques of event history analysis, developed to handle this challenging issues arising from the scale and complexity of lending longitudinal observational databases (LOD) . Typically, the self-controlled case series (SCCS) method provides insightful analysis in single repeated response events and the multivairate frailty method helps explain the high-dimentional heterogeneity and eliminate bias in latent confounding factors.

A Critical Evaluation of Large Language Models for Forecasting: Challenges and Opportunities

Claudio Antonini, CM, Inc.

Abstract: Using Large Language Models (LLMs) for forecasting raises a number of issues that cannot be found in more established quantitative methods. Traditional forecasting approaches, such as regression analysis, employ well-defined methodologies with es-

tablished testing procedures. For instance, regression analysis involves pre-application of tests to assess the suitability of time series data for the chosen method. Conversely, with LLMs, there is a lack of rigorous suitability assessment, sporadic arbitrary outputs, and limited repeatability. This presentation will examine two LLM forecasting applications. The first will be an inflation forecasting model proposed by a Federal Reserve Bank. We will highlight features of the LLM and its methodology that raise concerns about the validity of the results. Here, a comparison will be drawn between the LLM's output and the conclusions a skilled human analyst might reach using the same data. The second case study will showcase a successful application of LLMs, in this case, to forecast the likelihood and impact of global risk events. The LLM's results will be compared to surveys conducted annually to about a thousand individuals by a large-scale global organization. The discussion will highlight the limitations of the current survey approach, particularly regarding the consistency of results, despite their significant efforts in data collection and analysis. This comparison will underscore the potential advantages of LLMs in specific forecasting scenarios.

Innovative Statistical Modeling of Biomedical Big Data (IS-27)

Chair: Victor Hugo Lachos Davila

Proposer: Victor Hugo Lachos Davila, Fernanda Schumacher, University of Connecticut

Room: McHugh 205

Presenters: Fernanda Schumacher; Daniela Oliveira; Jiwon Park

Challenges in Assessing Biological Aging in People with Multiple Sclerosis Based on Epigenetic Clocks

Fernanda Lang Schumacher, The Ohio State University

Abstract: Multiple sclerosis (MS) is a complex autoimmune disease that affects the central nervous system, leading to various neurological symptoms and functional impairments. Recent research has explored the concept of biological aging in individuals with MS, aiming to understand the accelerated aging processes observed in this population. Among the several biomarkers of aging mechanisms that have been studied in humans, the study of epigenetic modifications represents a reversible mechanism in regulating the function of the genome without altering the underlying DNA sequence and has been linked to aging through several factors, including alterations in DNA methylation (DNAm). Specifically, patterns of cytosine-phosphate-guanine (CpG) site methylation are used as a reliable biomarker to predict chronological age, and epigenetic "clock" algorithms have been developed using statistical and machine learning methods to predict chronological age, biological age, and health outcomes in the general population and in specific disease conditions, based on variable selection. Since the first proposed epigenetic clock model in 2011, multiple epigenetic clocks have been reported with increasing accuracy, precision, and broader application prospects in aging research. Currently, a clinical study is underway at OSU aiming to test the hypothesis that aging biomarkers are more advanced in MS patients when compared to chronological age- and sex-matched healthy control. However, several epigenetic clocks have been implemented to measure epigenetic age and lifespan, and it is not clear how precise these algorithms are in predicting biological age in subjects with MS or how they relate to MS outcome measures. Given the variability in the CpG sites used for each clock, in addition to differences in additional clinical information considered, divergent conclusions might be reached based on different clocks, especially for subjects with MS. Additional challenges in analysis result from handling samples processed in batches, and from the interest in multivariate outcomes. In this talk, we will showcase and discuss such issues from a statistical point of view.

Emlmlasso Algorithm for Fixed Effects Selection in Linear Mixed Models with High-Dimensional and Correlated Genomic Data

Daniela Oliveira, Federal University of Sao Joao Del Rei

Abstract: The EM algorithm is a popular tool for maximum likelihood estimation but has not been used much for high-dimensional regularization problems in linear mixed-effects models. This study presents the EMLMLasso algorithm, which integrates the EM algorithm with the widely used and efficient R package glmnet, allowing for Lasso variable selection for fixed effects in such models. We assess its performance, comparing it to two existing algorithms from soft-

ware R called glmmLasso and splmm. Our findings, based on simulations and real data applications, illustrate our approach's robustness and effectiveness, even when the number of predictors surpasses the number of observations. Particularly, the EMLM-Lasso algorithm consistently outperforms glmmLasso and splmm across most scenarios. Additionally, our method is versatile and straightforward to implement, with the possibility to incorporate ridge and elastic net penalties in linear mixed-effects models.

A Robust Pleiotropic Analysis under Composite Null Hypothesis Exploring Shared Genetic Loci between Correlated Lipid Traits

Jiwon Park, Johns Hopkins University

Abstract: With the burgeoning interest in pleiotropy, where a single genetic variant affects multiple traits, the PLACO method was proposed to identify pleiotropic variants between two case-control traits, inclusive of sample overlap scenarios. We introduce the modified PLACO method, a novel scalable statistical approach based on GWAS summary statistics data for enhanced detection of pleiotropic variants across correlated quantitative or qualitative traits. By testing the composite null hypothesis that a variant is linked to at most one trait, the modified PLACO effectively controls type 1 errors and increases detection power for pleiotropy, especially in highly correlated traits. Applied to lipid traits— triglyceride and HDL levels—it unveils shared genetic regions overlooked by conventional methods, later validated by larger datasets. This demonstrates its ability to discover novel associations in traits often missed due to small sample sizes, later validated by larger datasets. This study highlights modified PLACO's potential for discovering novel genetic associations and offers a robust framework for pleiotropy analysis of two traits, regardless of their correlation or sample overlap.

Running The Gamut in Survival Analysis: Four Recent Results From Four Different Subfields (IS-38)

Chair: Jackson Lautier

Proposer: Jackson Lautier, Bentley University

Room: McHugh 206

Presenters: Ted Westling; Ying Chen; Sy Han (steven) Chiou; Jackson Lautier

Inference for Treatment-Specific Survival Curves Using Machine Learning

Ted Westling, University of Massachusetts Amherst

Abstract: In the absence of data from a randomized trial, researchers may aim to use observational data to draw causal inference about the effect of a treatment on a time-to-event outcome. In this context, interest often focuses on the treatment-specific survival curves, that is, the survival curves were the population under study to be assigned to receive the treatment or not. Under certain conditions, including that all confounders of the treatment-outcome relationship are observed, the treatment-specific survival curve can be identified with a covariate-adjusted survival curve. In this article, we propose a novel cross-fitted doubly-robust estimator that incorporates data-adaptive (e.g. machine learning) estimators of the conditional survival functions. We establish conditions on the nuisance estimators under which our estimator is consistent and asymptotically linear, both pointwise and uniformly in time. We also propose a novel ensemble learner for combining multiple candidate estimators of the conditional survival estimators. Notably, our methods and results accommodate events occurring in discrete or continuous time, or an arbitrary mix of the two. We investigate the practical performance of our methods using numerical studies and an application to the effect of a surgical treatment to prevent metastases of parotid carcinoma on mortality.

Regularized Estimation Methods for The Semiparametric AFT Model under Informative Sampling

Ying Chen, Harvard Chan school of public health

Abstract: In survival analysis, the semiparametric accelerated failure time (AFT) model is an attractive alternative to the Cox proportional hazards model, as it avoids the need for the PH assumption and provides a favorable interpretation of the covariate effect. However, the semiparametric AFT model is not widely used due to the lack of an efficient and reliable estimation algorithm. In this work, we develop two computationally effective regularized estimation methods for the semiparametric AFT model under informative sampling: the regularized least-square (LS) estimator and the regularized Gehan's weighted rank-based esti-

mator. Both two methods can accommodate sampling weights under non-randomized studies and are readily adapted with a broad class of penalty functions. Theoretical proprieties of the proposed estimator are provided. Simulation results indicate that the proposed methods efficiently reduce the sampling bias and achieve favorable variable selection performance under moderate sample sizes.

Regression Analysis of Bivariate Survival Data Using Pseudo-Observations

Sy Han Chiou, Southern Methodist University

Abstract: Copula models have become increasingly popular in various fields as they provide effective tools for modeling correlated responses. In modeling multivariate survival data, copula models offer flexibility by enabling users to specify both the marginal survival functions and the association structure between them. In this study, we consider a semiparametric transformation model to define the marginal survival functions and a conditional Archimedean copula to address the associations among different types of survival times. To expedite computation, we introduce pseudo-observations for both the marginal survival and association components and implement inference using generalized estimating equation techniques. Additionally, we explore variable selection and goodness-of-fit tests to aid in the selection of appropriate copula models. The effectiveness of our proposed methods is demonstrated through extensive simulations.

A Discrete-Time, Semi-Parametric Time-To-Event Model for Left-Truncated and Right-Censored Data

Jackson Lautier, Bentley University

Abstract: Insurance companies have large investment holdings in asset-backed securities (ABS). For the purposes of risk management, asset-liability management, and asset management generally, it is desirable to perform asset-level financial modeling of these ABS holdings. Any model calibration will rely upon empirical analysis of time-to-event data that is sampled from ABS trusts. This financial time-to-event observational data will be subject to discrete-time, left-truncation, and right-censoring, with a known, finite duration. This incomplete data combination has not received thorough study, however. In this paper, we propose a semi-parametric, discrete-time lifetime model that is attuned to left-truncated and right-censored data with a known, finite duration. We do not assume any forms for the left-truncation distribution, which offers more flexibility than the uniform assumptions of length-biased sampling. We then derive general stationary point theorems to maximize the likelihood function in both the cases of left-truncation only and left-truncation and right-censoring. Our results significantly simply a multiparametric constrained optimization problem into a single-parameter optimization problem. In the case of modeling lifetime data with a right-truncated geometric distribution, which is theoretically reasonable for payment time-to-event consumer loan data, we derive analytical results for the maximum likelihood estimates in both the cases of left-truncation and left-truncation with right-censoring. All theoretical results are illustrated with consumer auto loan data sampled from the Drive Auto Receivables Trust 2017-1 ABS bond.

Recent Development in Statistical Methodologies for Clinical Trials (IS-42)

Chair: Yiming Zhang

Proposer: Yiming Zhang, Center for Drug Evaluation and Research (CDER), FDA

Room: McHugh 301

Presenters: Peiran Liu; Cheng Huang; Eric Baron; Ling Zhu

Are The Tests Overpowered Or Underpowered? A Unified Solution to Correctly Specify Type I Errors in Design of Clinical Trials for Two Sample Proportions.

Peiran Liu, FDA

Abstract: As one of the most commonly used data types, methods in testing or designing a trial for binary endpoints from two independent populations are still being developed until recently. However, the power and the minimum required sample size comparisons between different tests may not be valid if their type I errors are not controlled at the same level. In this article, we unify all related testing procedures into a decision framework, including both frequentist and Bayesian methods. Sufficient conditions of the type I error attained at the boundary of hypotheses

are derived, which help reduce the magnitude of the exact calculations and lay out a foundation for developing computational algorithms to correctly specify the actual type I error. The efficient algorithms are thus proposed to calculate the cutoff value in a deterministic decision rule and the probability value in a randomized decision rule, such that the actual type I error is under but closest to, or equal to, the intended level, respectively. The algorithm may also be used to calculate the sample size to achieve the prespecified type I error and power. The usefulness of the proposed methodology is further demonstrated in the power calculation for designing superiority and noninferiority trials.

Bayesian Interim Analysis in Basket Trials

Cheng Huang, Vir Biotechnology

Abstract: Basket trials have captured much attention in recent years, as advances in health technology have opened up the possibility of classification of patients at the genomic level. Bayesian methods are particularly prevalent in basket trials as the hierarchical structure is adapted to basket trials to allow for information borrowing. In this talk, several existing Bayesian methods are extended to basket trials with treatment and concurrent control arms and continuous endpoints. Due to small sample sizes in basket trials, randomization often is not stratified on potentially strong predictors. Our methods not only allow adjustment for covariates but also basket-specific coefficients for the covariates. The performance of four Bayesian methods will be compared in both one-stage and two-stage designs.

Platform Trial Designs with Information Borrowing as An Efficient Drug Development Paradigm

Eric Baron, Servier Pharmaceuticals

Abstract: There has been a considerable increase in the use of platform trials due to their efficiency in concurrently testing multiple treatments against a shared control group with a single study. We consider an expansion platform trial setting with a common standard of care (SoC) and multiple experimental treatment arms with the potential of information sharing across the experimental treatment arms. The objective of the trial setting is to determine which, if any, of the experimental treatment arms are the most promising to graduate for potential future development. During the trial, interim analyses allow the discontinuation of any treatment arms that do not show meaningful improvement in comparison to the SoC based on monitoring rules. To enhance the efficiency of the interim analyses, this presentation applies a multi-source exchangeability modeling (MEM) approach to facilitate information sharing across possibly exchangeable arms. A simulation study will be presented to compare with methods with varying degrees of information borrowing.

Leveraging Historical Data via Propensity Score Matching: Optimizing The Selection of "Time Zero" for Controls to Minimize Bias

Ling Zhu, Vertex Pharmaceuticals

Abstract: The use of external controls in the study design of clinical trials has attracted increasing interest due to practical considerations and greater availability of Real World Data (RWD). However, while offering advantages over randomized clinical trial (RCT), such as reducing ethical concerns of exposing patients to suboptimal treatment, the use of external controls presents an increased risk of introducing bias compared to RCTs. In this talk, we review several types and sources of such bias and their mitigation strategies. Specifically, we examine the critical role of "Time Zero" specification in bias mitigation, and present two methods for establishing "Time Zero" and forming comparator groups: the prevalent new-user design and snapshot matching. The unique strengths of these two methods in mitigating bias will be evaluated.

Statistical Methods for High-Dimensional and Complex Data (IS-45)

Chair: Tong Wang

Proposer: Shuangge (Steven) Ma, Rong Li, Yale University

Room: McHugh 201

Presenters: Jiping Wang; Yuanxing Chen; Lijun Wang; Tong Wang

Local Clustering for Functional Data

Yuanxing Chen, Yale University

Abstract: In functional data analysis, unsupervised clustering has been extensively conducted and has important implications. In most of the existing functional clustering analyses, it is assumed that there is a single clustering structure across the whole domain of measurement (say, time interval). In some data analyses, for example, the analysis of normalized COVID-19 daily confirmed cases for the U.S. states, it is observed that functions can have different clustering patterns in different time subintervals. To tackle the lack of flexibility of the existing functional clustering techniques, we develop a local clustering approach, which can fully data-dependently identify subintervals, where, in different subintervals, functions have different clustering structures. This approach is built on the basis expansion technique and has a novel penalization form. It simultaneously achieves subinterval identification, clustering, and estimation. Its estimation and clustering consistency properties are rigorously established. In simulation, it significantly outperforms multiple competitors. In the analysis of the COVID-19 case trajectory data, it identifies sensible subintervals and clustering structures.

False Discovery Rate Control via Data Splitting for Clustering

Lijun Wang, Yale University

Abstract: Testing for differences in features between clusters in various applications often leads to inflated false positives when practitioners use the same dataset to identify clusters and then test features, an issue commonly known as "double dipping". To address this challenge, inspired by data-splitting strategies for controlling the false discovery rate (FDR) in supervised regressions (Dai et al., 2022), we present a novel method that applies data-splitting to control FDR while maintaining high power in unsupervised clustering. We first divide the dataset into two halves, then apply the conventional testing-after-clustering procedure to each half separately and combine the resulting test statistics to form a new statistic for each feature. The new statistic can help control the FDR due to its property of having a sampling distribution that is symmetric around zero for any null feature. To further enhance stability and power, we suggest multiple data splitting, which involves repeatedly splitting the data and combining results. Our proposed data-splitting methods are mathematically proven to asymptotically control FDR in Gaussian settings. Through extensive simulations and analyses of single-cell RNA sequencing (scRNA-seq) datasets, we demonstrate that the data-splitting methods are easy to implement, adaptable to existing single-cell data analysis pipelines, and often outperform other approaches when dealing with weak signals and high correlations among features.

Wasserstein Generative Regression

Tong Wang, Yale School of Public Health

Abstract: In this paper, we propose a new and unified approach for nonparametric regression and conditional distribution learning. Our approach simultaneously estimates a regression function and a conditional generator using a generative learning framework, where a conditional generator is a function that can generate samples from a conditional distribution. The main idea is to estimate a conditional generator that satisfies the constraint that it produces a good regression function estimator. We use deep neural networks to model the conditional generator. Our approach can handle problems with multivariate outcomes and covariates, and can be used to construct prediction intervals. We provide theoretical guarantees by deriving non-asymptotic error bounds and the distributional consistency of our approach under suitable assumptions. We also perform numerical experiments with simulated and real data to demonstrate the effectiveness and superiority of our approach over some existing approaches in various scenarios.

Recent Advances of Latent Variable Models in Education and Psychology (IS-59)

Chair: Xiaojing Wang

Proposer: Xiaojing Wang, University of Connecticut

Room: McHugh 305

Presenters: Betsy Mccoach; Sandip Sinharay; Xiaojing Wang

Using Latent Variable Models to Design Affective Instruments

D. Betsy Mccoach, University of Connecticut

Abstract: Latent variable modeling plays an important role in the design and evaluation of affective instru-

ments. Traditionally, developers of Likert scaled affective instruments have used traditional factor analysis techniques, and have often failed to treat the Likert Scaled data as ordinal. More recently, ordinal factor analysis techniques have become more widespread; however, relatively few applied researchers make use of item response theory models, even though IRT models have the huge advantage of emphasizing the importance of considering item difficulty and provide explicit information about the ability of the scale to provide information to differentiate scores at different ability levels. However, the parameters from ordinal response models can easily be transformed into IRT parameters, allowing researchers consider their data from both a more traditional factor analytic perspective and from an IRT perspective. In this talk, I will highlight the utility of ordinal factor models and IRT models for the affective instrument development process.

Detection of Fraudulent Behavior on Educational Tests Using Latent-Variable Models

Sandip Sinharay, Educational Testing Service

Abstract: Fraud on educational tests is a pervasive concern in various contexts ranging from elementary school achievement tests to post-secondary training and professional credentialing tests. Widespread fraud has also inspired a growing body of research on approaches that utilize latent-variable models for detecting fraudulent behavior on educational tests. The presentation will start with a brief review of some recent approaches for detecting test fraud using latent-variable models. Two new methodological approaches—one each based on Bayesian decision theory and model-fit assessment—will then be described for detecting fraudulent behavior using latent-variable models. Simulated and real data will be used to demonstrate the usefulness of the approaches.

Modeling Latent Trajectory of Ability in Dynamic Item Response Theory Models

Xiaojing Wang, University of Connecticut

Abstract: Item Response Theory (IRT) models play a key role in measurement testing. In the computerized testing scenarios, we will observe a time series of item responses for an individual throughout the entire study period. Integrating time as a factor offers insights into learning how the individual's ability is changing over the time. Leveraging the flexibility of Gaussian process and its simplicity of parametric computation, we have put forward a nonparametric IRT model to learn the growth of one's ability. The complexity of the proposed model made Bayesian approach ideal in its analysis. To facilitate Bayesian computation, we have proposed a modified slice sampler to draw hyper-parameters in the joint posterior, resulting in a much stable sampler in practice. In addition, we have introduced the concepts of Bayesian surrogate residuals to access the goodness of our model fit, which can be used to assess the performance of IRT model in general. In this paper, we have conducted several simulation studies to validate the efficacy of our approaches and applied the proposed model into a real testing dataset.

Student Paper 5 (S5)

Chair: TBD

Organizer: Neil Spencer

Room: McHugh 306

Functional Factor Modeling of Brain Connectivity

Kyle Stanley, The Pennsylvania State University

Co-authors: Matthew Reimherr, Nicole A. Lazar

Abstract: Many fMRI analyses examine functional connectivity, or statistical dependencies among remote brain regions. Yet popular methods for studying whole-brain functional connectivity often yield results that are difficult to interpret. Factor analysis offers a natural framework in which to study such dependencies, particularly given its emphasis on interpretability. However, multivariate factor models break down when applied to functional and spatiotemporal data, like fMRI. We present a factor model for discretely-observed multidimensional functional data that is well-suited to the study of functional connectivity. Unlike classical factor models which decompose a multivariate observation into a "common" term that captures covariance between observed variables and an uncorrelated "idiosyncratic" term that captures variance unique to each observed variable, our model decomposes a functional observation into two uncorrelated components: a "global" term that captures long-range dependencies and a "local" term that captures

short-range dependencies. We show that if the global covariance is smooth with finite rank and the local covariance is banded with potentially infinite rank, then this decomposition is identifiable. Under these conditions, recovery of the global covariance amounts to rank-constrained matrix completion, which we exploit to formulate consistent loading estimators. We study these estimators, and their more interpretable post-processed counterparts, through simulations, then use our approach to uncover a rich covariance structure in a collection of resting-state fMRI scans.

Calf-Sbm: A Covariate-Assisted Latent Factor Stochastic Block Model

Sydney Louit, University of Connecticut

Co-authors: Panpan Zhang, Evan Clark, Alexander Gelbard, Niketna Vivek

Abstract: We propose a novel network generative model extended from the standard stochastic block model by concurrently utilizing observed node-level information and accounting for network-enabled nodal heterogeneity. The proposed model is so called covariate-assisted latent factor stochastic block model (CALF-SBM). The inference for the proposed model is done in a fully Bayesian framework. The primary application of CALF-SBM in the present research is focused on community detection, where a model-selection-based approach is employed to estimate the number of communities which is practically assumed unknown. To assess the performance of CALF-SBM, an extensive simulation study is carried out, including comparisons with multiple classical and modern network clustering algorithms. Lastly, the paper presents two real data applications, respectively based on an extremely new network data demonstrating collaborative relationships of otolaryngologists in the United States and a traditional aviation network data containing information about direct flights between airports in the United States and Canada

Doubly Non-Central Beta Matrix Factorization for Stable Dimensionality Reduction of Bounded Support Matrix Data

Anjali Albert, UMass Amherst

Co-authors: Aaron Schein, Patrick Flaherty

Abstract: We consider the problem of developing interpretable and computationally efficient matrix decomposition methods for matrices whose entries have bounded support. Such matrices are found in large-scale DNA methylation studies and many other settings. Our approach decomposes the data matrix into a Tucker representation wherein the number of columns in the constituent factor matrices is not constrained. We derive a computationally efficient sampling algorithm to solve for the Tucker decomposition. We evaluate the performance of our method using three criteria: predictibility, computability, and stability. Empirical results show that our method has similar performance as other state-of-the-art approaches in terms of held-out prediction and computational complexity, but has significantly better performance in terms of stability to changes in hyper-parameters. The improved stability results in higher confidence in the results in applications where the constituent factors are used to generate and test scientific hypotheses such as DNA methylation analysis of cancer samples.

Distributed Tensor PCA with Heterogeneous Data

Wenbo Jing, New York University

Co-authors: Yichen Zhang, Elynn Chen and Xi Chen

Abstract: As tensors become widespread in modern data analysis, Tucker low-rank Principal Component Analysis (PCA) serves as an important tool for dimension reduction on tensor datasets. Motivated by large-scale tensors that are often dispersed across different locations, this paper studies the problem of tensor PCA in a distributed setting where multiple tensors are prohibited from being pooled together. We first propose a distributed framework for aggregating the principal components of homogeneous tensors, and then generalize it to a heterogeneous setting where the tensors are generated from different underlying models that share certain common principal components. As an extension, we propose a transfer learning algorithm, where the aim is to improve the estimation accuracy for a target tensor by transferring knowledge from a related source tensor. Statistical guarantees are established for all proposed methods, demonstrating that our distributed methods achieve a sharp rate by correctly aggregating the shared information across the different tensors, with a moderate communication cost. We also provide distributed inference methods for constructing confidence regions using our estimators. Simulations and real data analyses further verify the advantages of our proposed distributed methods, especially for heterogeneous tensors.

Likelihood-Based Inference for Assessing Biases in Multi-Item Measurement of A Latent

Trait

Zeling He, Emory University

Co-authors: Felicia Goldstein

Abstract: Multi-item scales used to measure an underlying health state might unintentionally have biases when applied to diverse populations. Such disparities can be uniform across all items or be item-specific. Our motivating example is the 10-item Functional Assessment Questionnaire (FAQ), which measures instrumental activities of daily living and is used in the differential diagnosis among normal aging, mild cognitive impairment (MCI), and dementia. Despite its widespread use in neurophysiology, the FAQ is self-reported which raises concerns that the items in FAQ might be biased for different sociodemographic groups. Traditional approaches to investigate such disparities rely on heuristic methods or allow only limited investigation. To address these limitations, we proposed a rigorous and computationally efficient likelihood-based approach to inference that fully incorporates a set of covariates in a differential item functioning (DIF) framework as an extension of item response theory. Simulation studies confirm that our proposed procedure yields valid inferences on the possibly complex DIF effects. We applied the proposed method to detect DIF of FAQ among participants diagnosed with MCI in the Uniform Data Set from the National Alzheimer's Coordinating Center. Our analysis shows that, after controlling for age, sex, and years of education, all items in FAQ presented biases regarding race. Compared with other subgroups of the same underlying impaired functioning level, black males are less likely to report impairment in FAQ.

Parallel Session 6 | 04:00 PM - 05:40 PM, May 23

Careers in Academia: A Panel Discussion by NESS Nextgen (IS-21)

Chair: Dr. Gregory Vaughan

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

Room: McHugh 102

Abstract: A career in academia can be very fulfilling and offers many exciting opportunities, but knowing where to start can be more complicated than it seems. This session, sponsored by the NESS NextGen committee, will feature a panel discussion among four faculty members from a variety of institutions who have recently been on the job market and are excited to share their experiences looking for a position and building a career. The panel will discuss a variety of topics including the different kinds of academic positions that are available for statisticians, what the application and interview process was like, the approaches they used to land a job, and what strategies they are using to grow in academia.

Panelists:

Dr. Reagan Mozer (Bentley University)

Dr. Stavroula Chrysanthopoulou (Brown University)

Dr. Debarghya Mukherjee (Boston University)

Dr. Duncan Clarke (Williams College)

Advanced Estimation Methods for Complex Data Structures in Medical Studies and Statistical Networks (IS-22)

Chair: Glen Laird

Proposer: Ce Yang, Vertex Pharmaceuticals Inc.

Room: McHugh 205

Presenters: Zhuoran Wei; Ce Yang; Jialu Wang; Sai Ma

Generalized Estimating Equations for Hearing Loss Data with Specified Correlation Structures

Zhuoran Wei, Harvard T.H. Chan School of Public Health

Abstract: Due to the nature of pure-tone audiometry test, hearing loss data often has a complicated correlation structure. Generalized estimating equation (GEE) is commonly used to investigate the association between exposures and hearing loss, because it is robust to misspecification of the correlation matrix. However, this robustness typically entails a moderate loss of estimation efficiency in finite samples. This paper proposes to model the correlation coefficients and use second-order generalized estimating equations to estimate the correlation parameters. In simulation studies, we assessed the finite sample performance of our proposed method and compared it with other methods, such as GEE with independent, exchangeable and unstructured correlation structures. Our method achieves an efficiency gain which is larger for the coefficients of the covariates corresponding to the within-cluster variation (e.g., ear- level covariates) than the coefficients of cluster-level covariates. The efficiency gain is also more pronounced when the within-cluster correlations are moderate to strong, or when comparing to GEE with an unstructured correlation structure. As a real-world example, we applied the proposed method to data from the Audiology Assessment Arm of the Conservation of Hearing Study, and studied the association between a dietary adherence score and hearing loss.

Soft Classification and Regression Analysis of Audiometric Phenotypes of Age-Related Hearing Loss

Ce Yang, Vertex Pharmaceuticals Inc

Abstract: Age-related hearing loss has a complex etiology. Researchers have made efforts to classify relevant audiometric phenotypes, aiming to enhance medical interventions and improve hearing health. We leveraged existing pattern analyses of age-related hearing loss and implemented the phenotype classification via quadratic discriminant analysis. We herein propose a method for analyzing the exposure effects on the soft classification probabilities of the phenotypes via estimating equations. Under reasonable assumptions, the estimating equations are unbiased and lead to consistent estimators. The resulting estimator had good finite sample performances in simulation studies. As an illustrative example, we applied our proposed methods to assess the association between a dietary intake

pattern, assessed as adherence scores for the Dietary Approaches to Stop Hypertension diet calculated using validated food-frequency questionnaires, and audiometric phenotypes (older-normal, metabolic, sensory, and metabolic plus sensory), determined based on data obtained in the Nurses' Health Study II Conservation of Hearing Study, the Audiology Assessment Arm. Our findings suggested that participants with a more healthful dietary pattern were less likely to develop the metabolic plus sensory phenotype of age-related hearing loss.

A/B Testing in Network Data with Covariate-Adaptive Randomization

Jialu Wang, Vertex Pharmaceuticals Inc

Abstract: People linked together through a network often tend to have similar behaviors. This phenomenon is usually known as network interaction. Their covariates are often correlated with their outcomes as well. Therefore, one should incorporate both the covariates and the network information in a carefully designed randomization to improve the estimation of the average treatment effect (ATE) in network data. In this talk, we introduce a new adaptive design to balance both the network and the covariates. We show that the imbalance measures with respect to the covariates and the network are Op(1). We also demonstrate the relationships between the improved balances and the increased efficiency in terms of the mean square error (MSE). Numerical studies demonstrate the advanced performance of the proposed design regarding the greater comparability of the treatment groups and the reduction of MSE for estimating the ATE.

Novel Statistical Approaches to Complex Data Structures (IS-28)

Chair: Ofer Harel

Proposer: Jung Wun Lee, Department of Biostatistics, Harvard University

Room: McHugh 206

Presenters: Jung Wun Lee; Mila Sun; Benjamin Stockton; Lucas Da Cunha Godoy

Sensitivity Analysis for Nonignorable Missing Values in Blended Analysis Framework: A Study on The Effect of Bariatric Surgery via Electronic Health Records

Jung Wun Lee, Department of Biostatistics, Harvard University

Abstract: This paper establishes a series of sensitivity analyses to investigate the impact of missing values in the electronic health records (EHR) that are possibly missing not at random (MNAR). EHRs have gained tremendous interest due to their cost-effectiveness, but their employment for research involves numerous challenges, such as selection bias due to missing data. The blended analysis has been suggested to overcome such challenges, which decomposes the data provenance into a sequence of sub-mechanisms and uses a combination of inverse-probability weighting (IPW) and multiple imputation (MI) under missing at random assumption (MAR). In this paper, we expand the blended analysis under the MNAR assumption and present a sensitivity analysis framework to investigate the effect of MNAR missing values on the analysis results. We illustrate the performance of my proposed framework via numerical studies and conclude with strategies for interpreting the results of sensitivity analyses. In addition, we present an application of our framework to the DURABLE data set, an EHR from a study examining long-term outcomes of patients who underwent bariatric surgery.

Estimating Weighted Quantile Treatment Effects with Missing Outcome Data by Double Sampling

Shuo Mila Sun, Harvard T.H. Chan School of Public Health

Abstract: Causal weighted quantile treatment effects (WQTEs) complement standard mean-focused causal contrasts when interest lies at the tails of the counterfactual distribution. However, existing methods for estimating and inferring causal WQTEs assume complete data on all relevant factors, which is often not the case in practice, particularly when the data are not collected for research purposes, such as electronic health records (EHRs) and disease registries. Furthermore, these data may be particularly susceptible to the outcome data being missing-not-at-random (MNAR). This paper proposes to use double-sampling, through which the otherwise missing data are ascertained on a sub-sample of study units, as a strategy to mitigate bias due to MNAR data in estimating causal WQTEs. With the additional data, we present

identifying conditions that do not require missingness assumptions in the original data. We then propose a novel inverse-probability weighted estimator and derive its asymptotic properties, both pointwise at specific quantiles and uniformly across quantiles over some compact subset of (0,1), allowing the propensity score and double-sampling probabilities to be estimated. For practical inference, we develop a bootstrap method that can be used for both pointwise and uniform inference. A simulation study is conducted to examine the finite sample performance of the proposed estimators. We illustrate the proposed method using EHR data examining the relative effects of two bariatric surgery procedures on BMI loss three years post-surgery.

A Novel Imputation Method for Incomplete Angular Time Series with An Application to Pm2.5 Air Pollution Analysis

Benjamin Stockton, University of Connecticut

Abstract: Air pollution and associated meteorological conditions are typically collected and reported at regular intervals by monitoring stations. The data produced by these monitoring stations can be incomplete due to technical/mechanical errors, systemic errors (recording only once every 3 hours rather than hourly), or many other potential complications. In this talk/paper, we develop a novel imputation method by imposing an autoregressive structure on the projected normal distribution to model wind direction observations. The imputations can then used in a multiple imputation scheme to create several completed data sets and several corresponding fitted models with Rubin's rules or MCMC posterior stacking to combine the estimates. The proposed method was validated using simulation studies based on autoregressive regression models for a simulated PM2.5 response with wind direction and speed as predictors. We used our proposed imputation methods to model hourly PM2.5 data and wind direction and speed data collected at eight sites in Connecticut from January 2018 to December 2018.

Beyond Traditional Disease Mapping: Hausdorff-Gaussian Process for Spatiotemporal Analysis of Tuberculosis Incidence

Lucas Da Cunha Godoy, University of Connecticut

Abstract: Tuberculosis (TB) remains a significant global health challenge, and Brazil exemplifies the complexities of controlling this infectious disease. Reliable estimates and forecasts of TB incidence rates are crucial to guide public health policies. This study focuses on the high-burden municipalities of Eastern Rio Grande do Sul, Brazil. We propose a novel spatiotemporal model based on the Hausdorff-Gaussian process to analyze TB incidence data. This model incorporates spatial dependence dictated by the Hausdorff distance, allowing it to "borrow strength" from neighboring municipalities and generate more reliable estimates, particularly for smaller areas. Our analysis has two primary goals. First, we aim to generate accurate TB incidence estimates by incorporating municipality-specific characteristics through covariates and a spatiotemporal random effect. The model delivers trustworthy expected incidence rates, consequently allowing for calculating standardized incidence ratios (SIRs). Second, our model offers predictive capabilities, forecasting TB incidence rates one year ahead to support proactive public health planning. We demonstrate our model's effectiveness and competitive performance against other specialized areal data models. The insights gained from this study can guide policymakers in developing effective TB control strategies.

Exploring Heterogeneity in Treatment Effects (IS-30)

Chair: Qi Zhang

Proposer: Qi Zhang, University of New Hampshire

Room: McHugh 301

Presenters: Ron Coury; Xinyuan Chen; Michael Lingzhi Li; Sarah Robertson

Exploring Heterogeneity in Treatment Effects

Ron Coury, University of New Hampshire

Abstract: Heterogeneous treatment effect (HTE) estimation has importance in many applications, such as determining a drug's efficacy on a patient's outcome over time. Machine learning methods known as random forests have gained popularity for this task when the exact time of occurrence is known. In many practical situations, however, the outcome may only fall in an interval of known times, e.g., the onset of a disease occurring between two checkups. In this talk

we propose a new forest-based method that builds on a conditional inference framework and provides HTE estimation for interval-censored data.

A Bayesian Machine Learning Approach for Estimating Heterogeneous Survivor Causal Effects: Applications to A Critical Care Trial

Xinyuan Chen, Mississippi State University

Abstract: Assessing heterogeneity in the effects of treatments has become increasingly popular in the field of causal inference and carries important implications for clinical decision-making. While extensive literature exists for studying treatment effect heterogeneity when outcomes are fully observed, there has been limited development in tools for estimating heterogeneous causal effects when patient-centered outcomes are truncated by a terminal event, such as death. Due to mortality occurring during study follow-up, the outcomes of interest are unobservable, undefined, or not fully observed for many participants, in which case principal stratification is an appealing framework to draw valid causal conclusions. Motivated by the Acute Respiratory Distress Syndrome Network (ARDSNetwork) ARDS respiratory management (ARMA) trial, we developed a flexible Bayesian machine learning approach to estimate the average causal effect and heterogeneous causal effects among the always-survivors stratum when clinical outcomes are subject to truncation. We adopted Bayesian additive regression trees (BART) to flexibly specify separate mean models for the potential outcomes and latent stratum membership. In the analysis of the ARMA trial, we found that the low tidal volume treatment had an overall benefit for participants sustaining acute lung injuries on the outcome of time to returning home, but substantial heterogeneity in treatment effects among the always-survivors, driven most strongly by biologic sex and the alveolar-arterial oxygen gradient at baseline (a physiologic measure of lung function and source of hypoxemia). These findings illustrate how the proposed methodology could guide the prognostic enrichment of future trials in the field.

Statistical Performance Guarantee for Subgroup Identification with Generic Machine Learning

Michael Lingzhi Li, Harvard Business School

Abstract: Across a wide array of disciplines, many researchers use machine learning (ML) algorithms to identify a subgroup of individuals who are likely to benefit from a treatment the most ("exceptional responders") or those who are harmed by it. A common approach to this subgroup identification problem consists of two steps. First, researchers estimate the conditional average treatment effect (CATE) using an ML algorithm. Next, they use the estimated CATE to select those individuals who are predicted to be most affected by the treatment, either positively or negatively. Unfortunately, CATE estimates are often biased and noisy. In addition, utilizing the same data to both identify a subgroup and estimate its group average treatment effect results in a multiple testing problem. To address these challenges, we develop uniform confidence bands for estimation of the group average treatment effect sorted by generic ML algorithm (GATES). Using these uniform confidence bands, researchers can identify, with a statistical guarantee, a subgroup whose GATES exceeds a certain effect size, regardless of how this effect size is chosen. The validity of the proposed methodology depends solely on randomization of treatment and random sampling of units. Importantly, our method does not require modeling assumptions and avoids a computationally intensive resampling procedure. A simulation study shows that the proposed uniform confidence bands are reasonably informative and have an appropriate empirical coverage even when the sample size is as small as 100. We analyze a clinical trial of late-stage prostate cancer and find a relatively large proportion of exceptional responders.

Examining Heterogeneity When Extending Causal Inferences to A New Population

Sarah Robertson, Harvard T.H. Chan School of Public Health

Abstract: Selection of participants into a randomized trial may result in trial-participants having a different distribution of treatment effect modifiers compared to the target population of interest. When this occurs, the average treatment effect estimate from the trial does not apply to the target population. Instead, generalizability or transportability methods that require adjusting for a large number of covariates to ensure that the trial and target population are conditionally exchangeable can be used to allow the estimation of treatment effects that apply to the target population. Still, heterogeneity needs to be examined after transporting inferences to the target population, particularly when strong effect modifiers render the average treatment effect less relevant for guiding treatment decisions. When that is the case, we propose estimating subgroup-specific treatment effects or the conditional

average treatment effect (CATE) as a function of key effect modifiers may be more useful. We illustrate the methods using data from the Coronary Artery Surgery Study (CASS) to estimate subgroup-specific effects given history of myocardial infarction and baseline ejection fraction value in the target population of all trial-eligible patients with stable ischemic heart disease.

Artificial Intelligence and Machine Learning Application in Pharmaceutical Statistics (IS-40)

Chair: Xiaofei Bai

Proposer: Xiaofei Bai, Servier

Room: McHugh 201

Presenters: Nan Miles Xi; Yu Deng; Jinchun Zhang; Reuben Retnam

Understanding The Rare Inflammatory Disease Using Large Language Models and Social Media Data

Nan Miles Xi, Loyola University Chicago

Abstract: Sarcoidosis is a rare inflammatory disease characterized by the formation of granulomas in various organs. The disease presents diagnostic and treatment challenges due to its diverse manifestations and unpredictable nature. In this study, we employed a Large Language Model (LLM) to analyze sarcoidosis-related discussions on the social media platform Reddit. Our findings underscore the efficacy of LLMs in accurately identifying sarcoidosis-related content. We discovered a wide array of symptoms reported by patients, with fatigue, swollen lymph nodes, and shortness of breath as the most prevalent. Prednisone was the most prescribed medication, while infliximab showed the highest effectiveness in improving prognoses. Notably, our analysis revealed disparities in prognosis based on age and gender, with women and younger patients experiencing good and polarized outcomes, respectively. Furthermore, unsupervised clustering identified three distinct patient subgroups (phenotypes) with unique symptom profiles, prognostic outcomes, and demographic distributions. Finally, sentiment analysis revealed a moderate negative impact on patients' mental health post-diagnosis, particularly among women and younger individuals. Our study represents the first application of LLMs to understand sarcoidosis through social media data. It contributes to understanding the disease by providing data-driven insights into its manifestations, treatments, prognoses, and impact on patients' lives. Our findings have direct implications for improving personalized treatment strategies and enhancing the quality of care for individuals living with sarcoidosis.

Developing Large Language Models for Adverse Drug Event Detection in Tweets

Yu Deng, abbvie

Abstract: Post-marketing adverse drug events (ADE) identification is one of the most important phases of drug safety surveillance. Social media data such as Tweets contains rich patient and medication information and could potentially accelerate drug surveillance research. However, such information is usually locked in the text in social media data, making it difficult to be employed by traditional statistical approaches. In recent years, transformer-based large language models (LLMs) have shown promise in many medical-related tasks. In this study, we developed and fine-tuned six LLMs to perform ADE classification in Twitter data. We further evaluate the feature importance using SHAP plot. Several fine-tuned LLMs show good performance in ADE classification. The top features from SHAP feature importance reveals interesting predictors for ADE. Our study shows the potential of using LLMs in ADE prediction.

Covariate-Adjusted Value-Guided Subgroup Identification via Boosting

Jinchun Zhang, Merck & Co.,

Abstract: It is widely recognized that treatment effects could differ across subgroups of patients. Subgroup analysis, which assesses such heterogeneity, provides valuable information in developing personalized therapies. There has been extensive research developing novel statistical methods for subgroup identification. The recent contribution is a value-guided subgroup identification method that directly maximizes treatment benefit at the subgroup level for survival outcome, rather than relying on individual treatment effect estimation. In this paper, we first completed this framework by illustrating its application to continuous and binary outcomes. More importantly, we extended the original framework to

account for the prognostic effects and named this new method Covariate-Adjusted Value-guided subgroup identification via boosting (CAVboost). The original method directly used the outcome to formulate the value function for subgroup identification. Since the outcome can further be decomposed as prognostic effects and treatment effects, specifying the prognostic effects as the covariates of a model for the outcome can single out the treatment effects and improve the power to detect them across subgroups. Our proposed CAVboost was based on this key idea. It used a covariate-adjusted treatment effect estimator, instead of the outcome itself, to formulate the value function for subgroup identification. CAVboost estimates the treatment effect by using covariates to account for the prognostic effects, which mimics the idea of using covariates in an ANCOVA estimator. We showed that CAVboost could effectively improve the subgroup identification capability for both continuous and binary outcomes.

Automatic Speech Recognition Based Measures for Speech Intelligibility Assessment in Parkinson's Disease

Reuben Retnam, Takeda Pharmaceuticals

Abstract: The challenges associated with Parkinson's Disease include cognitive and linguistic impairments, often including speech impairment. A key marker of speech impairment is speech intelligibility. Previous work has demonstrated that uncertainty-based features derived from automated speech recognition systems are correlated with intelligibility. We demonstrate the generation of intelligibility-related features from neural networks such as OpenAI's Whisper, and derive an analysis pipeline that allows these complex features to successfully discriminate between speakers with Parkinson's disease and healthy controls.

Understanding Change in Dynamic and Evolving Networks (IS-41)

Chair: Daniel L Sussman

Proposer: Daniel L Sussman, Boston University

Room: McHugh 202

Presenters: Shiwen Yang; Ankan Ganguly; Tianyi Chen; Joshua Loyal

Attractor-Based Coevolving Dot Product Random Graph Model: Decoding Polarizing/Flocking Behavior in Dynamic Network

Shiwen Yang, Boston University

Abstract: We introduce the attractor-based coevolving dot product random graph model (ABCDPRGM) to analyze time-series network data manifesting polarizing or flocking behavior. Graphs are generated based on latent positions under the random dot product graph regime. We assign group membership to each node. When evolving through time, the latent position of each node will change based on its current position and two attractors, which are defined to be the centers of the latent positions of all of its neighbors who share its group membership or who have different group membership than it. Parameters are assigned to the attractors to quantify the amount of influence that the attractors have on the trajectory of the latent position of each node. We developed estimators for the parameters, demonstrated their consistency, and established convergence rates under specific assumptions. Through the ABCDPRGM, we provided a novel framework for quantifying and understanding the underlying forces influencing the polarizing or flocking behaviors in dynamic network data.

Mean-Field and Graphon Limits of Latent Position Particle Systems with Dynamic Random Networks

Ankan Ganguly, Boston University

Abstract: Consider an opinion dynamics model in which n agents are each equipped with a latent opinion (represented by a Euclidean vector). As the agents update their opinions, they are influenced by their own previous opinions as well as the average opinions of all neighboring agents in a certain time-varying random network. At any given time, two agents are connected by an edge of the network with a probability that depends on their latent opinions as well as the existence or non-existence of an edge between the agents at the previous time. We are interested in the limiting behavior of this model as n increases to infinity. We show that under suitable conditions, this model has a mean-field limit which can be characterized explicitly. Using this characterization, we provide a full description of the asymptotic distribution and conditional structure of the latent opinions and network of a k-agent random sample taken from

a population of n agents in the limit as $n \to \infty$. From this, we can derive two hydrodynamic limits: the first describes the average behavior of agents in the system, and the second describes the average conditional behavior of a particular agent's neighbors and non-neighbors given the agent's latent opinion. We finish with a characterization of the graphon limit of the random network and a multigraphon limit of the entire network trajectory.

Euclidean Mirrors and First-Order Changepoints in Network Time Series

Tianyi Chen, Johns Hopkins University

Abstract: We describe a model for a class of network time series whose evolution is governed by an underlying stochastic process, known as the latent position process, in which network evolution can be represented in Euclidean space by a curve, called the Euclidean mirror. We define the notion of a first-order changepoint for a time series of networks, and construct a family of latent position process networks with underlying first-order changepoints. We prove that a spectral estimate of the associated Euclidean mirror localizes these changepoints, even when the graph distribution evolves continuously, but at a rate that changes. Simulated and real data examples on organoid networks show that this localization identifies empirically significant shifts in network evolution.

Fast Variational Inference of Latent Space Models for Dynamic Networks Using Bayesian P-Splines

Joshua Loyal, Florida State University

Abstract: Latent space models (LSMs) are often used to analyze dynamic (time-varying) networks that evolve in continuous time. Existing approaches to Bayesian inference for these models rely on Markov chain Monte Carlo algorithms, which cannot handle modern large-scale networks. To overcome this limitation, we introduce a new prior for continuous-time LSMs based on Bayesian P-splines that allows the posterior to adapt to the dimension of the latent space and the temporal variation in each latent position. We propose a stochastic variational inference algorithm to estimate the model parameters. We use stochastic optimization to subsample both dyads and observed time points to design a fast algorithm that is linear in the number of edges in the dynamic network. Furthermore, we establish non-asymptotic error bounds for point estimates derived from the variational posterior. To our knowledge, this is the first such result for Bayesian estimators of continuous-time LSMs. Lastly, we use the method to analyze a large data set of international conflicts consisting of 4,456,095 relations from 2018 to 2022.

Journal of Data Science Invited Overview Lecture: Power Priors for Leveraging Historical Data: Looking Back and Looking Forward (IS-55)

Chair: Jun Yan

Proposer: Jun Yan, University of Connecticut

Room: McHugh 101

Presenters: Ming-hui Chen; Minge Xie; Chenguang Wang; Panpan Zhang

Power Priors for Leveraging Historical Data: Looking Back and Looking Forward

Ming-hui Chen, University of Connecticut

Abstract: Historical data or real-world data (RWD) are often available in clinical trials, genetics, health care, psychology, environmental health, engineering, economics, and business. The power priors have emerged as a useful class of informative priors for a variety of situations in which historical data are available. In this paper, an overview of the development of the power priors is provided. Various variations of the power priors are derived under a binomial regression model and a normal linear regression model. The development of software on the power priors is also briefly reviewed. Throughout this paper, the data from the Kociba study and the National Toxicology Program (NTP) study as well as the data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study are used to demonstrate the derivations of the power priors and their variations. A detailed analysis of the Kociba and NTP data and the ADNI data is carried out to further demonstrate the usefulness of the power priors and their variations in these real applications. Finally, the directions of future research on the power priors are discussed.

Decoding The Future: Navigating Statistical Challenges and Innovations in Gene Therapy Trials (IS-56)

Chair: CG Wang

Proposer: CG Wang, Regeneron

Room: McHugh 305

Presenters: Chenguang Wang; Sammi Tang; Bret Musser; Siyu Zhang

Developing Innovative Dose-Finding Algorithms for Gene Therapy Trials

Chenguang Wang, Regeneron

Abstract: In this presentation, we will introduce a novel dose-finding design for gene therapy trials. This design simultaneously evaluates toxicity, individual level gene functional activity, and population-level gene functional activity. The proposed design makes dose escalation decisions based on the posterior probability of next participant having appropriate gene functional activity and the posterior probability of average gene functional activity being in a target range. The presentation will highlight the importance in efficient dose-escalation while ensuring the safety of individual participant in gene therapy clinical trials.

Statistical Challenges and Innovation in The Era of Gene and Cell Therapies

Siyu Zhang, Vertex Pharmaceuticals

Abstract: Cell and gene therapy drug development is accelerating to an unprecedent level, and as of now, multiple cell and gene therapies were approved by FDA and EMA including both malignant and non-malignant settings. Unique features of cell and gene therapies in non-malignant setting typically include single treatment administration, potential substantial pre-treatment procedures and manufacturing process, long term follow up (i.e. for at least 15 years); as such, randomized and blinded clinical trial may not be feasible and may not be deemed ethical. Design of these clinical trials usually utilizes single arm trials in relatively small sample size with frequent interim analyses for safety monitoring and potentially claim early efficacy. The authors would like to discuss the statistical issues and share the experience in regulatory submissions, including constructing the appropriate estimand and missing data handling, analysis population, justification for the null hypothesis and type 1 error control.

Student Paper 6 (S6)

Chair: TBD

Organizer: Neil Spencer

Room: McHugh 306

Leveraging Sparsity in The Gaussian Linear Model for Improved Inference

Souhardya Sengupta, Department of Statistics, Harvard University

Co-authors: Lucas Janson

Abstract: We develop novel LASSO-based methods for coefficient testing (and, via inversion, confidence interval construction) and variable selection in the Gaussian linear model with n *ge* d. Our methods have the same finite-sample guarantees as their ubiquitous ordinary-least-squares-t-test-based analogues, yet have substantially higher power when the true coefficient vector is sparse. Our coefficient test, which we call the l-test, often empirically performs like the one-sided t-test (despite not being given any information about the sign), and in particular our confidence intervals are typically about 15% shorter than the standard t-test based intervals. The nature of the l-test directly provides a novel exact adjustment conditional on LASSO selection for post-selection inference, and subsequently applying standard multiple testing procedures to the resulting post-selection-valid p-values results in large power gains over existing variable selection methods. 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 our methods over existing work. In the course of developing these methods, we also derive novel properties of the LASSO in the Gaussian linear model that are of independent interest.

Identifying Genetic Variants for Obesity Incorporating Prior Insights: Quantile Regression

with Insight Fusion for Ultra-High Dimensional Data

Jiantong Wang, University of Cincinnati

Co-authors: Heng Lian, Yan Yu, Heping Zhang

Abstract: Obesity is of major public health concern. We strive to identify single nucleotide polymorphisms (SNPs) associated with obesity by proposing and applying a novel Quantile Regression with Insight Fusion (QRIF) approach that can integrate insights from established studies or domain knowledge to simultaneously select variables and modeling for ultra-high dimensional genetic data. In this work, we focus on high conditional quantiles of body mass index (BMI) as an obesity phenotype. The SNPs that we identify, such as rs2033236 in CDH9, rs964841 in TAFA2, rs1899689 in CADPS2, rs2186948 in KCTD1, rs1873691 in KCNMA1, and rs2569034 in SGCD, provide a comprehensive view of the underlying genetic risk factors for different levels of BMI. This may potentially pave the way for more precise and targeted strategies to prevent and treat obesity. The QRIF approach intends to balance the trade-off between the prior insights and the observed data while being robust to potential false information. We further establish the desirable asymptotic properties under the challenging non-differentiable check loss functions via Huber loss approximation and nonconvex SCAD penalty via local linear approximation. Finally, we develop an efficient algorithm for the QRIF approach. Our simulation studies further demonstrate its effectiveness.

Melody: Meta-Analysis of Microbiome Association Studies for Discovering Generalizable Microbial Signatures

Zhoujingpeng Wei, Universitry of Wisconsin-Madison

Co-authors: Guanhua Chen

Abstract: With the increasing number of microbiome studies available, the field of microbiome research anticipates establishing standards for best practices in data sharing, harmonization, and meta-analysis. These efforts aim to identify generalizable microbial signatures associated with covariates such as health outcomes, exposures, and molecular phenotypes. Standard meta-analysis protocols are not suitable for microbiome data due to its unique characteristics and strong batch effects. Here, we introduce Melody, a framework that generates, harmonizes, and combines summary association statistics across studies to accurately and robustly identify microbial signatures in meta-analysis. We demonstrate through extensive simulations that Melody outperforms existing meta-analysis strategies in distinguishing signatures from noise features. In the meta-analysis of five microbiome studies for colorectal cancer, we showcase the superior stability and prediction performance of Melody-identified signatures. In the meta-analysis of eight microbiome-metabolome association studies with 450 metabolites, we demonstrate that Melody is scalable to meta-omics association scans and more reliably identifies signatures than existing methods.

A Statistical Testing Framework for Difference and Similarity of Multiple Gaussian Graphical Models with Application to Proteomic Networks

Jiachen Chen, Boston University School of Public Health

Co-authors: Joanne Murabito and Kathryn Lunetta

Abstract: The Gaussian Graphical Model (GGM) is a statistical network approach that represents conditional dependencies among components. Most existing methods for estimating group differences in sparse GGMs only apply to comparisons between two groups, and the challenging problem of multiple testing across multiple GGMs persists. To address this, we propose a new two-step statistical framework that tests for structural differences and similarities among multiple groups with false discovery rate (FDR) controlled asymptotically at a desired level. Our work focuses on entry-wise comparisons of precision matrices across groups and proposes two asymptotically independent test statistics for estimating structural differences and similarities sequentially while adjusting for the correlated effects of entry-wise test statistics in the FDR control procedures. We show via simulations that the proposed framework outperforms existing methods under a range of graph structures and is a valuable tool for joint comparisons of multiple GGMs. We also illustrate our method through the detection of potential neuroinflammatory pathways in a proteomics dataset involving three apolipoprotein E genotype groups.

Target-Oriented Reference Construction for Supervised Cell Type Identification in Scrna-Seq

Xin Wei, Brown University

Co-authors: Hao Wu

Abstract: As Single-cell RNA sequencing becomes increasingly applied, particularly in large-scale population-level studies, the identification of cell types remains among the most crucial aspects in single-cell RNA-seq data analysis. The accuracy and efficiency make the supervised cell type identification method an ideal solution. We propose a widely applicable strategy: Target-Oriented Reference Construction (TORC), which is based on a two-round supervised learning algorithm. The method utilizes the information gained from the first round prediction to construct a target-oriented reference as the input for the second round prediction. This approach alleviates the distributional shift as well as the difference in cell-type composition between reference and target data. We primarily use multilayer perceptron (MLP) to demonstrate the effectiveness and practicality of TORC, and perform systematic benchmarking on simulated experiments and real data analyses, demonstrating consistent improvements in both within-study and cross-study predictions. Furthermore, we showcase that the reference built upon MLP from the first round prediction could further benefit other supervised methods.

Parallel Session 7 | 08:45 AM - 10:25 AM, May 24

External Data Borrowing for Drug Approval: Review and Experience Sharing (IS-18)

Chair: Lanju Zhang

Proposer: Lanju Zhang, Vertex Pharmaceuticals

Room: McHugh 101

Presenters: Yiyue Lou; Jianchang Lin; Tianyu Sun; Yang Song

Leveraging External Data in Drug Development: Current Landscape and Methodological Considerations

Yiyue Lou, Vertex Pharmaceuticals

Abstract: The integration of external data in drug development and regulatory submissions has emerged as a critical aspect of modern pharmaceutical research. This talk provides a comprehensive overview of the current landscape of utilizing external data to strengthen regulatory submissions, focusing on four key areas. First, we examine regulatory guidelines and circumstances of utilizing external control to support drug approval. Next, we scrutinize the evaluation criteria for external data sources. Furthermore, we elucidate a systematic approach to mitigate potential biases inherent in external control data. Lastly, we discuss design and analytical considerations pertinent to both single-arm trials with external controls and randomized controlled trials augmented by external data. By delving into these four dimensions, this talk aims to equip pharmaceutical researchers and regulatory professionals with valuable insights and methodologies to effectively leverage external data in drug development endeavors.

Incorporating External Real-World Data (Rwd) in Confirmatory Adaptive Design

Jianchang Lin, Takeda Pharmaceuticals

Abstract: Adaptive designs, such as adaptive group sequential designs (and the ones with additional adaptive features) or adaptive platform trials, have been quintessential efficient design strategies in trials of unmet medical needs, especially for generating evidence from global regions. Such designs allow interim decision making and making adjustment to study design when necessary, meanwhile maintaining study integrity and operating characteristics. However, driven by the heightened competitive landscape and the desire to bring effective treatment to patients faster, innovation in the already functional designs is still germane to further propel drug development to a more efficient path. One way to achieve this is by leveraging external real-world data (RWD) in the adaptive designs to support interim or final decision making. In this paper, we propose a novel framework of incorporating external RWD in adaptive design to improve interim and/or final analysis decision making. Within this framework, researchers can prespecify the decision process and choose the timing and amount of borrowing while maintaining objectivity and controlling of type I error. Simulation studies in various scenarios are provided to describe power, type I error, and other performance metrics for interim/final decision making. A case study in non-small cell lung cancer (NSCLC) is used for illustration on proposed design framework.

Assemble External Controls From Real-World Data for Single Arm Trials with Small Sample Size: Utilizing Multiple Entries

Tianyu Sun, Moderna

Abstract: Conducting randomized clinical trials for medications targeting rare diseases presents significant challenges, due to the scarcity of participants and ethical considerations. Under such circumstances, leveraging real-world data (RWD) to generate supporting evidence may be accepted by the regulatory agency. Constructing external control arm (ECA) from RWD for a single arm trial has been conducted occasionally. A complication in this design is that patients from RWD may be eligible at multiple time points. Most studies approach this by selecting one time point as the index date for ECA patients. Here, we propose a novel framework for designing externally controlled trials that permits the inclusion of ECA patients at various entry points. Accompanying this design, we make recommendations of statistical methods to account for measured confounders, limited sample size, within-subject correlation, and potential overdispersion inherent in count data. Furthermore, we present an idea for the blinding process for this type of study. We have conducted a series of simulations to assess the performance of the framework in terms of bias, type I error, and efficiency, as compared to the approach of selecting only one entry per ECA patient. The parameter setup was based on a hypothetical case inspired by a rare disease study. The

results indicate that a design which allows for multiple entries for ECA patients can lead to enhanced performance. It provides a controlled type I error, robustness against certain model misspecifications, and a moderate power improvement compared with selecting a single entry per ECA patient.

Job Hunting and Career Development in Statistics and Data Science (IS-23)

Chair: Yang Liu

Proposer: Yang Liu, Upstart Network, Inc.

Room: McHugh 102

Abstract: This session will be in the form of a panel discussion, featuring several speakers who work in Statistics and data science related fields to share their experience on job hunting and career development. Current speakers include senior data scientists and Statistician/Biostatisticians with hands-on experiences working in industry and government, as well as a veteran recruiter in the technology sector. Speakers will share valuable experience and advice on job searching, resume preparation and career development.

Panelists:

1) Leander Eberhard, Upstart Network, Inc. Email: leander.eberhard@upstart.com

2) Nathan Bick, DV01, Inc. Email: nathankbick1@gmail.com

3) Catherine Reynolds, Upstart Network, Inc. Email: catherine.reynolds@upstart.com

4) Daniel Chen, Hartford Steam Boiler. Email: dchen37@gmail.com

5) Yiming Zhang, FDA. Email: Yiming.Zhang@fda.hhs.gov

6) Susan Wang, Boehringer Ingelheim. Email: susan.wang@boehringer-ingelheim.com

Statistical Learning Incorporating Data Structures (IS-43)

Chair: Rong Li

Proposer: Shuangge (Steven) Ma, Rong Li, Yale University

Room: McHugh 201

Presenters: Haoyu Yang; Weijuan Liang; Jiping Wang; Rong Li

Estimation Strategies for Treatment and Spillover Effects under Network Interference via Balancing

Haoyu Yang, Harvard T.H. Chan School of Public Health

Abstract: This study introduces a novel approach for treatment and spillover effects estimation in observational studies on social networks with arbitrary interference. We propose the direct covariate balancing estimator which is robust to model misspecification and avoids the extreme weights to gain the finite sample efficiency. To the best of our knowledge, this is the first attempt to adopt direct covariate balancing strategies in causal effect estimation under interference. We further improve the balancing estimator with the regrouping strategy to accommodate the limited sample sizes and vertex heterogeneity. We also advocate balancing the individual covariates as well as the network embeddings to safeguard the complexity of the data-generating process. Both theoretical and numerical justifications are established. Through the analysis of a real social experiment, the proposed method reveals the heterogeneity of conditional treatment effects, which sheds some light on the complexity of networked experiments.

Hierarchical False Discovery Rate Control for High-Dimensional Survival Analysis with Interactions

Weijuan Liang, School of Public Health, Yale University

Abstract: With the development of data collection techniques, analysis with a survival response and high-dimensional covariates has become routine. Here we consider an interaction model, which includes a set of low-dimensional covariates, a set of high-dimensional covariates, and their interactions. This model has been motivated by gene-environment (G-E) interaction analysis, where the E variables have a low dimen-

sion, and the G variables have a high dimension. For such a model, there has been extensive research on estimation and variable selection. Comparatively, inference studies with a valid false discovery rate (FDR) control have been very limited. The existing high-dimensional inference tools cannot be directly applied to interaction models, as interactions and main effects are not "equal". In this article, for high dimensional survival analysis with interactions, we model survival using the Accelerated Failure Time (AFT) model and adopt a "weighted least squares + debiased Lasso" approach for estimation and selection. A hierarchical FDR control approach is developed for inference and respect of the "main effects, interactions" hierarchy. The asymptotic distribution properties of the debiased Lasso estimators are rigorously established. Simulation demonstrates the satisfactory performance of the proposed approach, and the analysis of a breast cancer dataset further establishes its practical utility.

A Deep Neural Network-Based Approach for Network Analysis of Disease Clinical Treatment Measures via Mining Seer-Medicare Data

Jiping Wang, Yale University

Abstract: Extensive research has been conducted to analyze clinical treatment measures, bringing benefits for improved clinical resource management and a deeper understanding of diseases. Partly motivated by the successes of gene-centric human disease network (HDN) research, there has been growing interest in network analysis of clinical treatment measures. However, existing studies have been limited by neglecting zero-inflated data and making stringent model assumptions. In this study, we aim to extract insights from the SEER-Medicare data and construct HDNs for the number of in/outpatient visits. The proposed Deep Neural Network-based approach can accommodate a high proportion of zeros in the data and capture intricate conditional relationships between pairs of diseases. Simulation demonstrates the competitive performance of the proposed approach. In the analysis of SEER-Medicare data, we focus on patients diagnosed with different cancers, primarily in the context of cancer care. The interconnections, hubs, and network modules for different cancer populations are found to have sound implications.

Incorporating Prior Information in Gene Expression Network-Based Cancer Heterogeneity Analysis

Rong Li, Yale University

Abstract: Cancer is molecularly heterogeneous, with seemingly similar cancer patients having different molecular landscapes and accordingly different clinical behaviors. In recent studies, gene expression networks have been shown as more effective/informative for cancer heterogeneity analysis than some simpler measures. Gene interconnections can be classified as "direct" and "indirect", where the latter can be caused by shared genomic regulators (such as transcription factors, microRNAs, and other regulatory molecules). It has been suggested that incorporating the regulators of gene expressions in network analysis and focusing on the direct interconnections can lead to a deeper understanding of the more essential gene interconnections. Such analysis can be seriously challenged by the large number of parameters (jointly caused by network analysis, incorporation of regulators, and heterogeneity) and often weak signals. To effectively tackle this problem, we propose incorporating prior information contained in the published literature. A key challenge is that such prior information can be partial or even wrong. We develop a two-step procedure that can flexibly accommodate different levels of prior information quality. Simulation demonstrates the effectiveness of the proposed approach and its superiority over relevant competitors. In the analysis of a breast cancer dataset, findings different from the alternatives are made, and the identified sample subgroups have important clinical differences.

Advances in Latent Factorization Methods for Biomedical Data (IS-48)

Chair: Neil Spencer

Proposer: Jeffrey W Miller, Harvard TH Chan School of Public Health

Room: McHugh 202

Presenters: Phillip Nicol; Gemma Moran; Blake Hansen; Jenna Landy

Dimension Reduction for Single-Cell and Spatial Rna-Seq Using Generalized Bilinear Models

Phillip Nicol, Harvard Biostatistics

Abstract: Dimension reduction is a critical step in

the analysis of single-cell RNA-seq (scRNA-seq) data. The standard approach is to apply a transformation to the count matrix followed by principal component analysis (PCA). However, this approach can induce spurious heterogeneity and mask true biological variability. An alternative approach is to directly model the counts, but existing methods tend to be computationally intractable on large datasets and do not quantify uncertainty in the low-dimensional representation. To address these shortcomings, we develop scGBM, a novel method for model-based dimensionality reduction of scRNA-seq data. scGBM employs a scalable algorithm based on weighted low rank approximations to fit a Poisson bilinear model to datasets with millions of cells. Furthermore, scGBM quantifies the uncertainty in each cell's latent position and leverages these uncertainties to assess the confidence associated with a given cell clustering. Finally, we discuss potential extensions of scGBM to spatially resolved RNA-seq data by enforcing spatial smoothness in the estimated factors.

Spike-And-Slab Lasso Biclustering

Gemma Moran, Rutgers University

Abstract: Biclustering methods simultaneously group samples and their associated features. In this way, biclustering methods differ from traditional clustering methods, which utilize the entire set of features to distinguish groups of samples. Motivating applications for biclustering include genomics data, where the goal is to cluster patients or samples by their gene expression profiles; and recommender systems, which seek to group customers based on their product preferences. Biclusters of interest often manifest as rank-1 submatrices of the data matrix. This submatrix detection problem can be viewed as a factor analysis problem in which both the factors and loadings are sparse. In this paper, we propose a new biclustering method called Spike-and-Slab Lasso Biclustering (SSLB) which utilizes the Spike-and-Slab Lasso of Rockova and George (2018) to find such a sparse factorization of the data matrix. SSLB also incorporates an Indian Buffet Process prior to automatically choose the number of biclusters. Many biclustering methods make assumptions about the size of the latent biclusters; either assuming that the biclusters are all of the same size, or that the biclusters are very large or very small. In contrast, SSLB can adapt to find biclusters which have a continuum of sizes. SSLB is implemented via a fast EM algorithm with a variational step. In a variety of simulation settings, SSLB outperforms other biclustering methods. We apply SSLB to both a microarray dataset and a single-cell RNA-sequencing dataset and highlight that SSLB can recover biologically meaningful structures in the data. The SSLB software is available as an R/C++ package at https://github.com/gemoran/SSLB.

Fast Variational Inference for Bayesian Factor Analysis in Single and Multi-Study Settings

Blake Hansen, Brown University

Abstract: Factors models are commonly used to analyze high-dimensional data in both single-study and multi-study settings. Bayesian inference for such models relies on Markov Chain Monte Carlo (MCMC) methods, which scale poorly as the number of studies, observations, or measured variables increase. To address this issue, we propose new variational inference algorithms to approximate the posterior distribution of Bayesian latent factor models using the multiplicative gamma process shrinkage prior. The proposed algorithms provide fast approximate inference at a fraction of the time and memory of MCMC-based implementations while maintaining comparable accuracy in characterizing the data covariance matrix. We conduct extensive simulations to evaluate our proposed algorithms and show their utility in estimating the model for high-dimensional multi-study gene expression data in ovarian cancers. Overall, our proposed approaches enable more efficient and scalable inference for factor models, facilitating their use in high-dimensional settings. An R package VIMSFA implementing our methods is available on GitHub (github.com/blhansen/VI-MSFA).

Mutational Signatures in Practice with The Bayesnmf R Package

Jenna Landy, Harvard University

Abstract: Mutational signatures analysis is an unsupervised approach that models mutational patterns produced by various biological processes. Understanding the combination of mutational processes that are active in a cancer genome can help understand cancer development and guide questions about cancer subtyping, prognosis, and treatment. This talk introduces the biological motivation for mutational signatures analysis, the computational approaches used, and open research areas in the field. We introduce the bayesNMF R package, which implements a computationally efficient Bayesian NMF Gibbs sampler with learned rank, and present initial findings from simulation studies.

New Statistical Methods for Network Science (IS-52)

Chair: Krista Gile

Proposer: Krista Gile, University of Massachuestts, Amherst

Room: McHugh 205

Presenters: Daniel Sussman; Dongah Kim; Isabelle Beaudry; Amirhossein Alvandi

Causal Inference under Network Interference with Noise

Daniel Sussman, Boston University

Abstract: In this talk we consider the problem of estimating direct and indirect causal effect under a 4-exposure network interference model, where the network is observed with noise. We show that standard estimates are biased and proposed an estimate to reduce bias. We investigate these methods using a three contact networks collected at different points in time at the same school.

Bayesian Resolution of Discrepant Self-Reported Network Ties

Dongah Kim, University of Texas, Austin

Abstract: Most social network analysis assumes an objective network of shared social ties, typically measured as self-reports from research subjects. Although it is common for two parties to give discrepant reports on whether or not they are tied, there is no standard way to resolve such discrepancies in network data. Building on recent work, we develop a Bayesian model that leverages patterns of agreement among respondents across multiple relations, using flexible priors to allow for aberrant reporting behaviors. The model allows us to estimate error rates for individuals and to arrive at a probabilistic posterior estimate of the objective network. We demonstrate our method using data from the Food, Activity, Screens, and Teens (FAST) study, an investigation of social networks and health behavior among U.S. middle school students. The method discounts the reports of unreliable respondents in the FAST empirical data, and in simulations we find that our method performs better overall than the extreme conservative and extreme liberal methods commonly used to resolve discordant reports.

Distance Dependent Bayesian Community Detection

Isabelle Beaudry, Mount Holyoke College

Abstract: Community detection in networks seeks to identify to which cluster the nodes belong. Various traditional methods consider only the connections among the members of the population to make this classification. The rationale behind those methods is that highly connected nodes are more likely to belong to the same cluster than poorly connected ones. Although structure-based community detection algorithms may perform well under some circumstances, these methods ignore information included in many real-world networks, such as the characteristics of their members. For instance, nodal attributes may represent a person's characteristics, such as age, gender, and interests. The node attributes may affect the formation of clusters in a population. This phenomenon is well documented in the social science literature and is known as network homophily. Consequently, models leveraging link density and attribute homogeneity may help improve community detection. The main contribution of this work is to propose a community detection method that guarantees a lower link density across clusters than within clusters for any community pairs and also considers the similarity of the nodes with respect to their attributes. The inference is then performed under a Bayesian nonparametric framework. We assess the performance of the method through an extensive simulation study. Furthermore, we compare our results in real data with state-of-the-art community detection methodology for network data.

Modeling Social Networks Using Presence-Only Data by Augmenting Respondent-Driven Sampling

Amirhossein Alvandi, University of Massachusetts Amherst

Abstract: Respondent-Driven Sampling (RDS) is a widely used method for recruiting samples from hidden or hard-to-reach populations via social connections among population members. Traditional RDS captures connections through coupon-based recruitment; this paper enhances these data with a novel augmentation involving the distribution of tokens, allowing the exploration of otherwise missed social ties. We employ a variant of logistic regression, adapted from Ward et al. (2009), which utilizes an Expectation-Maximization (EM) algorithm to handle positive unlabeled data—observed ties are labeled, while non-observed ties remain unlabeled (NA). This model

specifically addresses unobserved social relationships not directly captured through token distribution. To further understand the underlying social structure, we model the population network using an Exponential Random Graph Model (ERGM), controlling for node mean degree and incorporating variables reflecting homophily levels and differential group activity. The performance of our estimator is assessed under various simulation settings, examining both the sampled nodes' subgraph and the complete population graph, and is applied to data collected among people who inject drugs in Kenya.

Recent Development in Spatial and Spatiotemporal Modeling (IS-54)

Chair: Yeongjin Gwon

Proposer: Yeongjin Gwon, University of Nebraska Medical Center

Room: McHugh 206

Presenters: Malcolm Itter; Rajarshi Mukherjee; Mary Lai Salvana

A Dynamical Spatio-Temporal Model to Predict Regional Forest Dynamics

Malcolm Itter, University of Massachusetts Amherst

Abstract: Models of forest dynamics are an essential tool to predict forest ecosystem responses to global change. These predictions are used to make management and policy decisions to maintain forest health, function, and ecosystem services. To maximize the utility of such predictions, contemporary models of forest dynamics must address the uncertainty in forest demographic processes under no-analog conditions. Dynamical spatio-temporal models (DSTMs) applied within an iterative, near-term forecasting framework are a particularly powerful tool in this setting given that they quantify and partition uncertainty in demographic models and noisy forest observations, propagate uncertainty to predictions of adaptive management outcomes, and refine predictions based on new monitoring data and improved ecological understanding. A major challenge to the application of DSTMs in applied forest ecology has been the lack of a scalable, theoretical model of forest dynamics that generates predictions at the stand level—the scale at which management decisions are made. We address this challenge by integrating a matrix projection model motivated by the well-known McKendrick-von Foerster partial differential equation for size-structured population dynamics within a Bayesian hierarchical DSTM informed by continuous forest inventory data. The model provides probabilistic predictions of species-specific demographic rates and changes in the size-species distribution of stands in response to climate, disturbance, and alternative management approaches. We apply the model framework to predict the long-term dynamics (60+ years) of stands under alternative management scenarios within the Penobscot Experimental Forest in Maine. We conclude with a discussion of how the DSTM can be scaled up to predict regional forest dynamics assimilating new inventory data as it becomes available.

On Statistical Inference under Spatiotemporal and Network Dependence

Rajarshi Mukherjee, Harvard T.H. Chan School of Public Health

Abstract: We discuss some recent work that aims to understand statistical issues and challenges while dealing with dependence that arises in many modern observational studies. Specifically, we will discuss a case study in spatiotemporal causal inference, extract the salient features of the challenges that arise in such contexts, and present some mathematical formalisms along with numerical simulations to sketch the contours of subtleties that arise while dealing with inference under dependence under such contexts.

Multi- and Mixed-Precision Computations for Spatial and Spatio-Temporal Statistics

Mary Lai Salvana, University of Connecticut

Abstract: Computational statistics has traditionally utilized double-precision (64-bit) data structures and full-precision operations, resulting in higher-than-necessary accuracy for certain applications. Recently, there has been a growing interest in exploring low-precision options that could reduce computational complexity while still achieving the required level of accuracy. This trend has been amplified by new hardware such as NVIDIA's Tensor Cores in their V100, A100, and H100 GPUs, which are optimized for mixed-precision computations, Intel CPUs with Deep Learning (DL) boost, Google Tensor Processing Units (TPUs), Field Programmable Gate Arrays

(FPGAs), ARM CPUs, and others. However, using lower precision may introduce numerical instabilities and accuracy issues. Nevertheless, some applications have shown robustness to low-precision computations, leading to new multi- and mixed-precision algorithms that balance accuracy and computational cost. To address this need, we introduce MPCR, a novel R package that supports three different precision types (16-, 32-, and 64-bit) and their combinations, along with its usage in commonly-used Frequentist/Bayesian statistical examples. The MPCR package is written in C++ and integrated into R through the Rcpp package, enabling highly optimized operations in various precisions. Moreover, we show how to leverage low precision computations for spatial and spatio-temporal statistics.

Recent Advances in Data-Driven Decision-Making and Generative Models (IS-60)

Chair: Henry Lam

Proposer: Henry Lam, Columbia University

Room: McHugh 305

Presenters: Hanyang Zhao; Arindam Roy Chowdhury

Contractive Diffusion Probabilistic Models

Hanyang Zhao, Columbia University

Abstract: Diffusion probabilistic models (DPMs) have emerged as a promising technology in generative modeling. The success of DPMs relies on two ingredients: time reversal of Markov diffusion processes and score matching. Most existing works implicitly assume that score matching is close to perfect, while this assumption is questionable. In view of possibly unguaranteed score matching, we propose a new criterion – the contraction of backward sampling in the design of DPMs, leading to a novel class of contractive DPMs (CDPMs). Our key insight lies in the illustration that the contraction in the backward process can narrow score-matching errors as well as discretization errors. Thus, our proposed CDPMs are theoretically robust to both sources of error. For practical concerns, we further show that CDPM does not need any retraining and can leverage pre-trained existing DPMs by a simple transformation. We corroborated our proposal by experiments on CIFAR-10 dataset: notably, CDPM shows the best performance among all known SDE-based DPMs, with pre-trained scores based on NCSN++.

Robustness Versus Statistical Efficiency: Superiority of Naive Optimization via Worst-Case Stochastic Dominance

Arindam Roy Chowdhury, Columbia University

Abstract: Many decision-making tasks give rise to optimization problems with uncertainty in the underlying parameter or probability distribution. This uncertainty can arise from statistical noises due to limited data, or non-stationary parameter shifts that are not predictable from the past. In the literature, various approaches ranging from (distributionally) robust optimization to regularization have been proposed, as attempts to obtain solutions that perform well statistically or robustly against model shifts. Contrary to common belief, we show that, in terms of the behavior of regret or optimality gap, a naive optimization formulation based on empirical or simple point estimate of the uncertain parameter is in a sense optimal, regardless of whether the uncertainty comes from statistical noises, unpredictable shifts, or both. Our superiority assertion for naive optimization is based on a new notion of worst-case stochastic dominance for the asymptotic regret, and is argued via what we call the "symmetry" of uncertainty that applies universally over statistical errors, adversarial errors, and their interactions. We show how our results apply to the comparisons among a wide range of existing data-driven and robust optimization formulations.

Theory and Methods on Statistical Learning and Machine Learning (IS-65)

Chair: TBD

Proposer: Kun Chen, University of Connecticut

Room: McHugh 301

Presenters: Ye Tian; Zeyuan Song; Stephen Bates; Awni Altabaa

Neyman-Pearson Multi-Class Classification via Cost-Sensitive Learning

Ye Tian, Department of Statistics, Columbia University

Abstract: Most existing classification methods aim to minimize the overall misclassification error rate. However, in applications such as loan default prediction, different types of errors can have varying consequences. To address this asymmetry issue, two popular paradigms have been developed: the Neyman-Pearson (NP) paradigm and the cost-sensitive (CS) paradigm. Previous studies on the NP paradigm have primarily focused on the binary case, while the multi-class NP problem poses a greater challenge due to its unknown feasibility. In this work, we tackle the multi-class NP problem by establishing a connection with the CS problem via strong duality and propose two algorithms. We extend the concept of NP oracle inequalities, crucial in binary classifications, to NP oracle properties in the multi-class context. Our algorithms satisfy these NP oracle properties under certain conditions. Furthermore, we develop practical algorithms to assess the feasibility and strong duality in multi-class NP problems, which can offer practitioners the landscape of a multi-class NP problem with various target error levels. Simulations and real data studies validate the effectiveness of our algorithms. To our knowledge, this is the first study to address the multi-class NP problem with theoretical guarantees. The proposed algorithms have been implemented in the R package "npcs", which is available on CRAN.

Learning Gaussian Graphical Models From Correlated Data

Zeyuan Song, Tufts Medical Center

Abstract: Gaussian Graphical Models (GGM) have been widely used in biomedical research to explore complex relationships between many variables. There are well established procedures to build GGMs from a sample of independent and identical distributed observations. However, many studies include clustered and longitudinal data that result in correlated observations and ignoring this correlation among observations can lead to inflated Type I error. In this paper, we propose a Bootstrap algorithm to infer GGM from correlated data. We use extensive simulations of correlated data from family-based studies to show that the Bootstrap method does not inflate the Type I error while retaining statistical power compared to alternative solutions. We apply our method to learn the GGM that represents complex relations between 47 Polygenic Risk Scores generated using genome-wide genotype data from a family-based study known as the Lo ng Life Family Study. By comparing it to the conventional methods that ignore within-cluster correlation, we show that our method controls the Type I error well in this real example.

Incentive-Theoretic Bayesian Inference for Collaborative Science

Stephen Bates, MIT

Abstract: Contemporary scientific research is a distributed, collaborative endeavor, carried out by teams of researchers, regulatory institutions, funding agencies, commercial partners, and scientific bodies, all interacting with each other and facing different incentives. To maintain scientific rigor, statistical methods should acknowledge this state of affairs. To this end, we study hypothesis testing when there is an agent (e.g., a researcher or a pharmaceutical company) with a private prior about an unknown parameter and a principal (e.g., a policymaker or regulator) who wishes to make decisions based on the parameter value. The agent chooses whether to run a statistical trial based on their private prior and then the result of the trial is used by the principal to reach a decision. We show how the principal can conduct statistical inference that leverages the information that is revealed by an agent's strategic behavior – their choice to run a trial or not. In particular, we show how the principal can design a policy to elicit partial information about the agent's private prior beliefs and use this to control the posterior probability of the null. One implication is a simple guideline for the choice of significance threshold in clinical trials: the type-I error level should be set to be strictly less than the cost of the trial divided by the firm's profit if the trial is successful.

On The Role of Information Structure in Reinforcement Learning for Partially-Observable Sequential Teams and Games

Awni Altabaa, Yale University, Department of Statistics & Data Science

Abstract: In a sequential decision-making problem, the information structure is the description of how events in the system occurring at different points in time affect each other. Classical models of reinforcement learning (e.g., MDPs, POMDPs, Dec-POMDPs, and POMGs) assume a very simple and highly regular information structure, while more general models like predictive state representations do not explicitly

model the information structure. By contrast, real-world sequential decision-making problems typically involve a complex and time-varying interdependence of system variables, requiring a rich and flexible representation of information structure. In this paper, we argue for the perspective that explicit representation of information structures is an important component of analyzing and solving reinforcement learning problems. We propose novel reinforcement learning models with an explicit representation of information structure, capturing classical models as special cases. We show that this leads to a richer analysis of sequential decision-making problems and enables more tailored algorithm design. In particular, we characterize the "complexity" of the observable dynamics of any sequential decision-making problem through a graph-theoretic analysis of the DAG representation of its information structure. The central quantity in this analysis is the minimal set of variables that d-separates the past observations from future observations. Furthermore, through constructing a generalization of predictive state representations, we propose tailored reinforcement learning algorithms and prove that the sample complexity is in part determined by the information structure. This recovers known tractability results and gives a novel perspective on reinforcement learning in general sequential decision-making problems, providing a systematic way of identifying new tractable classes of problems.

Student Paper 7 (S7)

Chair: TBD

Organizer: Neil Spencer

Room: McHugh 306

Causal Inference on Sequential Treatments via Tensor Completion

Chenyin Gao, North Carolina State University

Co-authors: Anru Zhang

Abstract: Marginal Structural Models (MSMs) are popular for causal inference of sequential treatments in longitudinal observational studies, which however are sensitive to model misspecification. To achieve flexible modeling, we envision the potential outcomes to form a three-dimensional tensor indexed by subject, time, and treatment regime and propose a tensorized history-restricted MSM. The semi-parametric tensor factor model allows us to leverage the underlying low-rank structure of the potential outcomes tensor and exploit the pre-treatment covariate information to recover the counterfactual outcomes. We incorporate the inverse probability of treatment weighting in the loss function for tensor completion to adjust for time-varying confounding. Theoretically, a non-asymptotic upper bound on the Frobenius norm error for the proposed estimator is provided. Empirically, simulation studies show that the proposed tensor completion approach outperforms the parametric HRMSM and existing matrix/tensor completion methods. Finally, we illustrate the practical utility of the proposed approach to study the effect of ventilation on organ dysfunction from the Medical Information Mart for Intensive Care database.

Studying Disease Reinfection Rates, Vaccine Efficacy and The Timing of Vaccine Rollout in The Context of Infectious Diseases

Elizabeth Amona, Virginia Commonwealth University

Co-authors: Ryad Ghanam

Abstract: The global landscape has undergone distinct waves of COVID-19 infections, compounded by the emergence of variants, thereby introducing additional complexities to the ongoing pandemic. This research uniquely explores the varied efficacy of existing vaccines and the pivotal role of vaccination timing in the context of COVID-19. Departing from conventional modeling, we introduce two models that account for the impact of vaccines on infections, reinfections, and deaths. We estimate model parameters under the Bayesian framework, specifically utilizing the Metropolis-Hastings Sampler. The study conducts data-driven scenario analyses for the State of Qatar, quantifying the potential duration during which the healthcare system could have been overwhelmed by an influx of new COVID-19 cases surpassing available hospital beds. Additionally, the research explores similarities in predictive probability distributions of cumulative infections, reinfections, and deaths, employing the Hellinger distance metric. Comparative analysis, utilizing the Bayes factor, underscores the plausibility of a model assuming a different susceptibility rate to reinfection, as opposed to assuming the same susceptibility rate for both infections and reinfections. Results highlight the adverse outcomes associated with delayed vaccination, emphasizing the efficacy of early vaccination in reducing infections, re-

infections, and deaths. Our research advocates prioritizing early vaccination as a key strategy in effectively combating future pandemics. This study contributes vital insights for evidence-based public health interventions, providing clarity on vaccination strategies and reinforcing preparedness for challenges posed by infectious diseases.

Refined Methods for Trial Sequential Analyses in Living Systematic Reviews

Yipeng Wang, University of Florida

Co-authors: Lifeng Lin

Abstract: A living systematic review (LSR) is an evolving approach that aims to provide continuous updates and real-time synthesis of evidence. Unlike traditional systematic reviews conducted at a specific time, LSRs incorporate new studies as they become available. Compared to the study collection and qualitative assessment in LSRs, few works are devoted to developing quantitative methods. The trial sequential analysis (TSA) is a well-known procedure to assess the adequacy of the available evidence based on the collected studies in a LSR. It uses trial sequential monitoring boundaries for assessing the efficacy of an intervention and futility boundaries for evaluating whether the intervention does not differ significantly from the control. Although TSAs have recently gained popularity, existing TSA methods have limitations stemming from their heavy reliance on interim analyses of randomized controlled trials, where individuals are often more homogeneous than those in meta-analyses. For random-effects meta-analyses, the normality-based methods can perform poorly when the number of studies is small. In such cases, the Hartung–Knapp–Sidik–Jonkman (HKSJ) method based on the t-statistic is more appropriate. This article introduces novel trial sequential methods based on the t-statistic of the cumulative meta-analysis. The proposed methods can prevent LSRs from being terminated prematurely, allowing for more robust evidence syntheses. Numerical studies show that the proposed methods are more reliable than the existing methods.

Spatio-Temporal Quasi-Experimental Methods for Rare Disease Outcomes: The Impact of Reformulated Gasoline on Childhood Hematologic Cancer

Sofia Vega, Harvard University

Co-authors: Dr. Rachel C. Nethery

Abstract: Although some pollutants emitted in vehicle exhaust, such as benzene, are known to cause leukemia in adults with high exposure levels, less is known about the relationship between traffic-related air pollution (TRAP) and childhood hematologic cancer. In the 1990s, the US EPA enacted the reformulated gasoline program in select areas of the US, which drastically reduced ambient TRAP in affected areas. This created an ideal quasi-experiment to study the effects of TRAP on childhood hematologic cancers. However, existing methods for quasi-experimental analyses can perform poorly when outcomes are rare and unstable, as with childhood cancer incidence. We develop Bayesian spatio-temporal matrix completion methods to conduct causal inference in quasi-experimental settings with rare outcomes. Selective information sharing across space and time enables stable estimation, and the Bayesian approach facilitates uncertainty quantification. We evaluate the methods through simulations and apply them to estimate the causal effects of TRAP on childhood leukemia and lymphoma.

Unobserved Heterogeneity in Threshold Regression Based on The Hitting Times of A Reflected Brownian Motion for Recurrent Hypoglycemia

Yingfa Xie, University of Connecticut

Co-authors: Haoda Fu, Yuan Huang, and Jun Yan

Abstract: Analyses of recurrent hypoglycemia are critical for effective treatment management of diabetic patients. Typically, within-subject dependency in such analyses captured through subject-level frailty. Recent research has utilized the first hitting times of a reflected Brownian motion to model recurrent hypoglycemia among diabetic patients. A close scrutiny of this approach shows that the frailties vary significantly among individuals, indicating notable heterogeneity. To address this, we propose a finite mixture model of the first hitting time distribution of the reflected Brownian motion. This model allows for component-specific regression coefficients and frailty parameters, providing nuanced insights into how risk factors differently affect patient subgroups. We employ a Bayesian framework for inference, utilizing Markov chain Monte Carlo for estimation. Model selection is conducted using the Deviance Information Criterion and the Logarithm of the Pseudo-Marginal Likelihood. The effectiveness of these criteria is assessed through simulation studies. This approach enhances our understanding of variability in risk factor impacts across patient subgroups, offering a more tailored analysis of recurrent hypoglycemic events. Application to recur-

rent hypoglycemia modeling revealed two subgroups with different risk levels, reflected in their volatilities. Bayesian model comparison criteria prefer the model with component specific regression coefficients in volatilities. The subgroup with lower volatility exhibits a higher variance and, hence, a greater level of heterogeneity.

Parallel Session 8 | 01:30 PM - 03:10 PM, May 24

Advanced Design Techniques for Data Science (IS-13)

Chair: HaiYing Wang

Proposer: Weng Kee Wong, UCLA

Room: McHugh 201

Presenters: Qiong Zhang; Nicholas A Rios; Lulu Kang; Jing Wang

Statistical Designs for Network A/B Testing

Qiong Zhang, Clemson University

Abstract: A/B testing is a common controlled experiment approach used to compare two versions of internet-based products. IT companies often conduct A/B testing on their users who are connected in a social network. The users' responses could be related to the network connection, leading to two typical assumptions: network interference and network correlated outcomes. I will discuss the general problems of network A/B testing design and their relationship with graph cut objectives under the two assumptions. Further, I will show the asymptotic distributions of graph cut objectives to enable rerandomization algorithms for the design of network A/B testing.

Graphical Methods for Order-Of-Addition Experiments

Nicholas Rios, George Mason University

Abstract: In an order-of-addition (OofA) experiment, the order in which several components are added to a system influences a response. Although much research has been done on optimal OofA experiments, existing methodologies typically assume that all orders are possible. However, in many practical examples, there are directed constraints on the pairwise order of components, making some of the orders infeasible. These constraints can be represented by a directed acyclic graph (DAG). The goal of the OofA experiment is to find an optimal order, which is equivalent to finding an optimal topological sort of the DAG. A multiplicative algorithm is used to identify approximate optimal designs for an arbitrary DAG. Simulated annealing (SA) is proposed as a method to identify efficient exact designs. It is shown that the SA designs have high efficiency relative to the approximate optimal designs. A general procedure is proposed to search for the optimal order on a DAG given the results of an OofA experiment using two popular models. Applications to job scheduling are shown.

Optimal Kernel Learning for Gaussian Process Models with High-Dimensional Input

Lulu Kang, University of Massachusetts Amherst

Abstract: Many computer simulation models in engineering and scientific domains involve a large number of input variables, which can result in high computational cost and low prediction accuracy for the Gaussian process (GP) regression model. However, some simulation models may only be significantly influenced by a small subset of the input variables, referred to as the "active variables". Identifying these active variables can help researchers overcome the two limitations of the GP model and gain a better understanding of the simulated system. To achieve this goal, we propose an approximation of the covariance function of the original GP model involving all the input variables. The approximation is through a convex combination of kernel functions whose input variables are low-dimensional subsets of the complete input variables. To determine the optimal approximation, we develop an iterative algorithm based on the Fedorov-Wynn algorithm from the optimal design literature. We also incorporate the effect heredity principle while selecting the active input variables, which ensures sparsity. Through several examples, we have shown the proposed method outperforms some alternative approaches in correctly identifying the active input variables.

Optimal Subsampling for Transfer-Leanring

Jing Wang, University of Connecticut

Abstract: Transfer-learning is an emerging field in recent years. Subsampling can be considered as a data selection method for transfer-learning to obtain better performances. However, optimal subsampling with potential model misspecification has not been full investigate which limits the usage of subsampling algorithms in transfer-learning. In this paper, we develop subsampling algorithms with potential mean shifts, which connects subsampling under misspecified models with data selection for transfer-learning algorithms. Theoretical analysis implies that the performances of transfer-leanring estimators are determined by model biases and variances. Therefore, we propose two different subsampling strateges, one to reduces model

biases and the other reduces variances due to subsampling. We also propose two approaches to combine the two sampling strateges to further improve the performances of transfer-learning estimators. Non-asymptotic bounds of the proposed estimators are proved. Numerical experiments justify the usage of the proposed transfer-leanring algorithms.

Innovative Statistical Modeling with Applications in Epidemiology and Public Health (IS-20)

Chair: Priya Kohli

Proposer: Priya Kohli, Connecticut College

Room: McHugh 202

Presenters: Elizabeth Upton; Mengyan Li; Yichi Zhang; Gregory Vaughn

Assessing The Conditional Dependencies of Post-Covid Symptoms Using Network Analysis

Elizabeth Upton, Williams College

Abstract: While most people with COVID-19 recover within a few weeks, many others develop lasting symptoms. The diversity of symptoms in patients suffering from Post-COVID Conditions (PCC) underscores the need to understand and quantify how these symptoms evolve and co-occur. Using data from a multinational, randomized, placebo-controlled trial, we analyze symptom diaries, assessing the severity of 27 targeted symptoms, completed by 2824 participants spanning a period of 12 to 72 weeks post initial COVID diagnosis. We construct a network composed of the 27 symptoms by employing a generalization of the Ising model and penalized regression techniques. We utilize a bootstrapping algorithm to capture variability in edge weights representing conditional dependencies between pairs of symptoms. The Walktrap community detection algorithm is used to group the symptoms into highly connected subgroups, and we analyze a variety of centrality statistics to identify core symptoms in the network. Furthermore, we perform permutation tests to determine differences in inferred networks based on participant strata.

Multi-Source Graph Synthesis (Mugs) for Pediatric Knowledge Graphs From Electronic Health Records

Mengyan Li, Bentley University

Abstract: The wealth of valuable real-world medical data found within Electronic Health Record (EHR) systems is particularly significant in the field of pediatrics, where conventional clinical studies face notably high barriers. However, constructing accurate knowledge graphs from pediatric EHR data is challenging due to its limited content density compared to EHR data for the general population. Additionally, knowledge graphs built from EHR data primarily covering adult patients may not suit the unique biomedical characteristics of pediatric patients. In this research, we introduce a graph transfer learning approach aimed at constructing precise pediatric knowledge graphs. We present MUlti-source Graph Synthesis (MUGS), an algorithm designed to create embeddings for pediatric EHR codes by leveraging information from three distinct sources: (1) pediatric EHR data, (2) EHR data from the general population, and (3) existing hierarchical medical ontology knowledge shared across different patient populations. We break down these code embeddings into shared and unshared components, facilitating the adaptive and robust capture of varying levels of heterogeneity across different medical sites through meticulous hyperparameter tuning. We assessed the quality of these code embeddings in recognizing established relationships among pediatric codes, as curated from credible online sources, pediatric physicians, or GPT. Furthermore, we developed a web API for visualizing pediatric knowledge graphs generated using MUGS embeddings and devised a phenotyping algorithm to identify patients with characteristics similar to a given profile, with a specific focus on pediatric pulmonary hypertension (PH). The MUGS-generated embeddings demonstrated resilience against negative transfer and exhibited superior performance across all three tasks when compared to pediatric-only approaches, multi-site pooling, and semantic-based methods. MUGS embeddings open up new avenues for evidence-based pediatric research utilizing EHR data.

Addressing Post-Treatment Selection Bias in Comparative Effectiveness Research, Using Real-World Data and Simulation

Yichi Zhang, University of Rhode Island

Abstract: To examine methodologies that address imbalanced treatment switching and censoring, six

different approaches were evaluated under a comparative effectiveness framework: intention-to-treat, as-treated, intention-to-treat with censor-weighting, as-treated with censor-weighting, time-varying exposure, and time-varying exposure with censor-weighting. Marginal structural models were employed to address time-varying exposure, confounding, and possibly informative censoring in an administrative data set of adult patients who were hospitalized with acute coronary syndrome and treated with either clopidogrel or ticagrelor. The effectiveness endpoint was the occurrence of death, myocardial infarction, or stroke. These methodologies were then applied across simulated data sets with varying frequencies of treatment switching and censoring. The findings suggest that implementing different approaches has an impact on the point estimate and interpretation of analyses, especially when censoring is highly unbalanced.

Racial Disparities in The Burden of Disease and Cost of Stroke

Greg Vaughan, Bentley University

Abstract: With the cost of healthcare an ever-growing concern, some argue for a shift to payment on the quality, efficiency, and effectiveness of care provided (value-based pricing), rather than the quantity of services (fee for service). However, implementation of a value-based model may have unintended consequences if racial or socio-economic contributors to disease are not accounted for. This study examines racial disparities in the health burden among Black and white patients at high risk for stroke and heart disease using data from the Reasons for Geographic and Racial Differences in Stroke (REGARDS) study. The finding of a higher burden of disease among underrepresented populations, even after accounting for socioeconomic factors, suggests a value-based approach could unintentionally lead to discriminatory pricing if not accounted for. Using the Institute for Clinical and Economic Review recommended costs for measured treatment health benefit, achieving equivalent health outcomes could cost over $100,000 more for Black stroke patients.

Drawing Causal Conclusions with Observational Data: Recent Theory and Methods (IS-32)

Chair: Ted Westling

Proposer: Ted Westling, University of Massachusetts Amherst

Room: McHugh 205

Presenters: Nicole Pashley; Rohit Bhattacharya; Jessica Young; Edward Kennedy

Instrumental Variable Methods for Factorial Experiments with Complex Treatment Uptake

Nicole E. Pashley, Rutgers University

Abstract: There is a well-established literature dealing with noncompliance in treatment-control designs within the potential-outcome framework for causal inference. However, generalizing to experiments with more than two treatment arms and noncompliance remains a challenge with limited exploration. The focus of this talk will be noncompliance in two-level factorial designs. The talk will discuss why this setting is so challenging, propose different assumptions to learn about relevant estimands, and explore identification and inference results under these assumptions.

Using Experimental Data To Evaluate Observational Causal Inference Methods: How, When, and Why

Rohit Bhattacharya, Williams College

Abstract: Rigorous empirical evaluation of observational causal inference methods is challenging. Unlike supervised learning problems which have ground-truth labels for evaluating predictive performance on a held-out test set, analogous causal estimation problems require ground-truth labels for counterfactual outcomes of an individual under multiple versions of the treatment, data that is generally impossible to measure. In this work, we build on a promising empirical evaluation strategy that simplifies evaluation design and uses real data: subsampling randomized controlled trials (RCTs) to create confounded observational datasets while using the average causal effects from the RCTs as ground-truth. This idea has appeared in several works like Hill (2011) and Zhang & Bareinboim (2021) and was recently formalized by Gentzel et al. (2021). Here, we present theory that clarifies why and how RCT subsampling algorithms

should be constrained in order to produce valid downstream empirical comparisons. In particular, we prove that previous subsampling algorithms can produce observational samples from which the causal effect is not identifiable, which makes recovery of the RCT ground-truth (and thus, an informative empirical comparison) impossible. To address this issue, we present a new RCT rejection sampling algorithm that appropriately constrains the subsampling such that the observed data distribution permits identification. In addition to this theoretical result, we highlight several finite data considerations for evaluation designers who plan to use RCT rejection sampling on their own datasets. As a proof of concept, we implement an example evaluation pipeline and walk through these finite data considerations using data from a real-world RCT—which we release publicly—consisting of approximately 70k observations and text data as high-dimensional covariates. Together, these contributions build towards a broader agenda of improved empirical evaluation for causal estimation.

Estimating Causal Effects of Generalized Time-Varying Treatment Strategies on Continuous Health Marker Outcomes in Electronic Health Records

Jessica G. Young, Harvard Medical School and Harvard Pilgrim Health Care Institute

Abstract: Clinical researchers are often interested in leveraging electronic health record (EHR) data to inform the comparative effects of different treatment strategies on a continuous health marker (e.g. weight gain, blood pressure, cholesterol). One example is the Medications and Weight Gain in PCORnet (MedWeight) Study which is leveraging EHR data to inform comparative effects of initiating - and then continuing what was initiated "to some degree" - different medications for a common disease on weight change. "To some degree" here refers to a continuation strategy that pragmatically allows short"treatment breaks" as well as stopping or switching given development of contraindications. The broader umbrella MedWeight study captures 5 separate comparative studies in adult and pediatric populations, respectively, of different antidepressants, antipsychotics, diabetes medications, antiseizure medications, and antihypertensives. Several interesting elements converge in MedWeight that more broadly apply to many EHR studies of the effects of time-varying treatment strategies on a continuous health marker including: 1) adjustment for time-varying covariates affected by past treatment is needed; 2) the particulars of the strategies of interest have "non-deterministic" and "dynamic" components relative to past measured covariates; 3) the outcome of interest is repeatedly measured with nonmonotonic missingness patterns; and 4) the outcome is undefined for an individual after death such that "go-to" causal effect notions are themselves undefined (a problem broadly known as "truncation by death"). In this talk I will motivate and describe the analytic approach that I structured for the 10 MedWeight clinical papers which attempts to jointly address the first three elements. If time, I will discuss planned future work that includes extensions to further address the fourth element for this type of study.

Minimax Optimal Estimation of Sample Average Causal Effects

Edward Kennedy, Carnegie Mellon University

Abstract: Sample average causal effects can be preferred over population effects when samples are not iid draws from a larger population, for example when there is dependence between units and/or no clear superpopulation. In standard root-n settings, a common refrain is that sample effects can be estimated more accurately than population effects. However, optimality has remained an open problem; in this paper we resolve the question of minimax optimality for sample effects. Surprisingly, we show that in the non-root-n regime sample effects can only be estimated less accurately than population effects. The main issue is that randomness in the covariates (coupled with some knowledge of the covariate density) can allow for less biased estimation of nuisance regressions; when conditioning on covariate values, as in sample effects, exploiting such randomness is infeasible. In addition to characterizing lower and upper bounds on the minimax rate for sample effects, we also give new results for estimating matrix inverses and regression reciprocals, which are of independent interest. We conclude with simulation experiments and applications in political science.

Innovative Statistical Methodologies for The Design of Patient-Centric Clinical Trials (IS-34)

Chair: Meizi Liu

Proposer: Meizi Liu, Takeda

Room: McHugh 206

Presenters: Yu-che Chuang; Miachael Kane; Rachael Liu; Xiaofei Bai

Dod-Pro-Bart: Dose Optimization Design Incorporating Patient-Reported Outcomes via Bayesian Additive Regression Trees

Yu-che Chung, Takeda Pharmaceuticals

Abstract: In drug development, dose optimization is crucial and challenging due to the inherent variability and exploratory nature in early phase trials. It requires careful evaluation of dose-response and toxicity to ensure that the treatment is effective and safe while maintaining an acceptable level of tolerability. In this work, we propose the Dose Optimization Design incorporating Patient-Reported Outcomes via Bayesian Additive Regression Trees (DOD-PRO-BART). It is an innovative approach that integrates patient-reported outcomes with clinician-reported toxicity and efficacy data, enabling a more personalized and patient-centered method in both dose escalation and dose randomization. Our simulation study illustrates that the proposed method can substantially improve the optimal dose selection by integrating patient-reported data along with clinician-reported toxicity and efficacy data.

A Multiple Indication Dose Optimization Design Using Multisource-Exchangeability

Michael Kane, Yale Unviersity

Abstract: Project Optimus, initiated by the U.S. FDA, aims to revamp traditional dose determination methods for oncology indications, which primarily depended on the maximum tolerated dose (MTD). This shift is motivated by the emergence of new classes of therapies with comparatively safer profiles than cytotoxic chemotherapies. These agents, which attempt to manipulate specific aspects of the tumor microenvironment or inhibit oncogenic signaling pathways, often lack monotonically increasing dose response curves. Moreover, pre-clinical models often hypothesize that benefits to patients are agnostic to the primary tumor's tissue of origin, yielding development strategies that span multiple indications. Although, introducing multiple doses in trials offers comprehensive data on safety and efficacy, it increases trial complexity and potential variability in outcomes. The extended evaluation for each dose might also prolong the entire trial requiring more resources. Bayesian borrowing offers a solution to these challenges by enabling information sharing across patient subgroups within a trial. This talk presents a dose-equivalent basket trial design, rooted in specific foundational assumptions regarding dose-response curves for different indications within a single trial. By leveraging information borrowing across comparable dose levels for various indications, our method ensures more precise (higher power and narrower confidence intervals) response rate estimates and better delineates the dose-response curve, aiding healthcare providers in balancing dose efficacy with potential adverse events.

Beats: Bayesian Hybrid Design with Flexible Sample Size Adaptation for Time-To-Event Endpoints

Rachael Liu, Takeda Pharmaceuticals

Abstract: As the roles of historical trials and real-world evidence in drug development have substantially increased, several approaches have been proposed to leverage external data and improve the design of clinical trials. While most of these approaches focus on methodology development for borrowing information during the analysis stage, there is a risk of inadequate or absent enrollment of concurrent control due to misspecification of heterogeneity from external data, which can result in unreliable estimates of treatment effect. In this study, we introduce a Bayesian hybrid design with flexible sample size adaptation (BEATS) that allows for adaptive borrowing of external data based on the level of heterogeneity to augment the control arm during both the design and interim analysis stages. Moreover, BEATS extends the Bayesian semiparametric meta-analytic predictive prior (BaSe-MAP) to incorporate time-to-event endpoints, enabling optimal borrowing performance. Initially, BEATS calibrates the expected sample size and initial randomization ratio based on heterogeneity among the external data. During the interim analysis, flexible sample size adaptation is performed to address conflicts between the concurrent and historical control, while also conducting futility analysis. At the final analysis, estimation is provided by incorporating the calibrated amount of external data. Therefore, our proposed design allows for an approximation of an ideal randomized controlled trial with an equal randomization ratio while controlling the size of the concurrent control to benefit patients and accelerate drug development. BEATS also offers optimal power and robust estimation through flexible sample size adaptation when conflicts arise between the concur-

rent control and external data.

Applying CHW Method to 2-In-1 Design: Gain Or Lose?

Xiaofei Bai, Servier

Abstract: The 2-in-1 design allows the possibility of seamlessly expanding a phase II study to confirmatory phase III study and controls type I error without multiplicity adjustment. In this talk, we applied the CHW method to the 2-in-1 design strategy, and compared it with the unweighted conventional test statistics. It shows that when the interim decision threshold is high enough, the CHW method is slightly more powerful. Otherwise, results based on the CHW method can be difficult to interpret when the estimated treatment effects differ notably between interim and final analysis, which may be avoided by using the conventional test statistic.

Recent Advances in Variational Inference (IS-35)

Chair: Mike Wojnowicz

Proposer: Mike Wojnowicz, Harvard University

Room: McHugh 101

Presenters: Michael Wojnowicz; Diana Cai; Manushi Welandawe; Tamara Broderick

Scalable Bayesian Multi-Sample Changepoint Modeling

Michael Wojnowicz, Harvard University

Abstract: Changepoint models detect abrupt changes in sequential data, and are used in a range of applications in genetics, economics, and biometrics. Bayesian changepoint models naturally quantify uncertainty about changepoint locations, but they are highly sensitive to the choice of prior on changepoint probabilities. A hierarchical Bayesian model can address this concern by borrowing information across multiple samples, however, Bayesian multi-sample changepoint models have intractable posteriors, and existing approximate inference methods do not easily scale up to large datasets. In this talk, we show how to leverage variational inference to obtain fast, closed-form inference for Bayesian multi-sample changepoint models. We present promising initial results on simulated data, and consider the problem of identifying copy number alterations in cancer biopsy samples with low tumor fractions.

Batch and Match: Black-Box Variational Inference with A Score-Based Divergence

Diana Cai, Flatiron Institute

Abstract: Most leading implementations of black-box variational inference (BBVI) are based on optimizing a stochastic evidence lower bound (ELBO). But such approaches to BBVI often converge slowly due to the high variance of their gradient estimates. In this talk, we present "batch and match" (BaM), an alternative approach to BBVI based on a score-based divergence. Notably, this score-based divergence can be optimized by a closed-form proximal update for Gaussian variational families with full covariance matrices. We analyze the convergence of BaM when the target distribution is Gaussian, and we prove that in the limit of infinite batch size the variational parameter updates converge exponentially quickly to the target mean and covariance. We also evaluate the performance of BaM on Gaussian and non-Gaussian target distributions that arise from posterior inference in hierarchical and deep generative models. In these experiments, we find that BaM typically converges in fewer (and sometimes significantly fewer) gradient evaluations than leading implementations of BBVI based on ELBO maximization. Finally, we conclude with a discussion of extensions to richer variational families.

A Framework to Enhance The Reliability and Detect Convergence of Black-Box Variational Inference

Manushi Welandawe, Boston University

Abstract: Black-box variational inference (BBVI) has rapidly gained popularity in machine learning and statistics as a versatile alternative to Markov chain Monte Carlo methods for approximate Bayesian inference. However, the reliability of stochastic optimization methods for BBVI remains challenging, requiring significant expertise and manual tuning for effective implementation. In response, we introduce Robust and Automated Black-box VI (RABVI), a framework designed to enhance the reliability of BBVI optimization. RABVI incorporates rigorously justified automa-

tion techniques with only a few intuitive tuning parameters, while also detecting inaccurate estimates of the optimal variational approximation. It dynamically adjusts the learning rate by identifying convergence, estimates the symmetrized Kullback-Leibler (KL) divergence between the current and optimal variational approximations, and introduces a novel optimization termination criterion. This criterion balances desired accuracy against computational cost by comparing the predicted relative decrease in KL divergence with the predicted computation required for convergence. We validate the robustness and accuracy of RABVI through comprehensive simulation studies and application to various real-world models and data examples. Furthermore, to improve the sensitivity of convergence detection for stochastic optimization within BBVI, particularly when employing flexible approximation families, we introduce a novel method that overcomes limitations of the usual potential scale reduction factor $\widehat{R}$. This approach ensures reliable estimates across a diverse range of statistical models, thus enhancing the utility of BBVI in addressing complex inferential challenges.

Black Box Variational Inference with A Deterministic Objective: Faster, More Accurate, and Even More Black Box

Tamara Broderick, Massachusetts Institute of Technology

Abstract: Automatic differentiation variational inference (ADVI) offers fast and easy-to-use posterior approximation in multiple modern probabilistic programming languages. However, its stochastic optimizer lacks clear convergence criteria and requires tuning parameters. Moreover, ADVI inherits the poor posterior uncertainty estimates of mean-field variational Bayes (MFVB). We introduce "deterministic ADVI" (DADVI) to address these issues. DADVI replaces the intractable MFVB objective with a fixed Monte Carlo approximation, a technique known in the stochastic optimization literature as the "sample average approximation" (SAA). By optimizing an approximate but deterministic objective, DADVI can use off-the-shelf second-order optimization, and, unlike standard mean-field ADVI, is amenable to more accurate posterior covariances via linear response (LR). In contrast to existing worst-case theory, we show that, on certain classes of common statistical problems, DADVI and the SAA can perform well with relatively few samples even in very high dimensions, though we also show that such favorable results cannot extend to variational approximations that are too expressive relative to mean-field ADVI. We show on a variety of real-world problems that DADVI reliably finds good solutions with default settings (unlike ADVI) and, together with LR covariances, is typically faster and more accurate than standard ADVI.

Recent Developments in The Analysis of Time-To-Event Data with Cured Fractions (IS-53)

Chair: Austin Menger, Ph.D.

Proposer: Austin Menger, Ph.D, Bridge Mental, Inc.

Room: McHugh 301

Presenters: Austin Menger; Md. Tuhin Sheikh; Shike Xu; Hongfei Li

Bayesian Modeling of Survival Data in The Presence of Competing Risks with Cure Fractions and Masked Causes

Austin Menger, Bridge Mental, Inc.

Abstract: In handling the presence of multiple competing risks, methods such as the multivariate failure times model, mixture model, subdistribution model (partially and fully specified), and the cause-specific hazard model have historically been used under the assumption that all individuals will either "fail" or are censored. However, in practice, there may be a group of individuals who are actually cured of a given cause (but not necessarily all causes). The proposed model addresses this issue of cured fractions, using a mixture cure rate model with cause-specific hazards for the non-cured survival time. To handle the common issue of masked causes, where cause of death is not known, we incorporate these individuals into the likelihood function directly. A Bayesian approach to inference is proposed, rooted in the augmentation of the joint posterior distribution, using baseline survival covariates to model non-cured survival time. A more informative Jeffrey's-type prior is used for the cure rate model coefficients to help address identifiability in the model, necessitating the proposal of a more efficient sampling algorithm to avoid direct calculation of the derivative of the Jeffreys-type prior. A variation of the DIC and C-index measures are developed to allow for cause-specific assessment of the utility of

the proposed methodology not dependent upon the latent structure of the masked causes and cure rate indicators. Findings are demonstrated empirically using prostate cancer diagnosis data from the National Cancer Institute's Surveillance, Epidemiology, and End Results Program.

Bayesian Joint Model for Longitudinal Biomarker and Competing Risks of Prostate Cancer with Cure Fraction, Accounting for Masked Causes

Md Tuhin Sheikh, Postdoctoral Associate, Department of Biostatistics, Yale University

Abstract: In medical studies, subjects who never develop a disease or are cured of it pose challenges, especially in competing risks survival data with partially masked causes. Motivated by the SELECT data, where prostate cancer (PC) can arise from observed (low-grade, high-grade) or masked causes (unknown-grade), and prostate-specific antigens (PSA) are collected longitudinally as a potential biomarker, we propose a Bayesian joint model with a mixed effects regression sub-model for longitudinal data and a competing risks double-regression promotion time cure rate sub-model incorporating the random effects for survival data. We employ an efficient Markov chain Monte Carlo (MCMC) sampling algorithm to carry out posterior computation. We develop a variation of the deviance information criterion for assessing the fit of the cause-specific survival data. We also propose a novel cause-specific concordance (C)-index by utilizing augmented latent tumor carcinogenic cell counts within each iteration of MCMC sampling to quantify the discriminatory and predictive performance of the cause-specific survival models. The necessary theory and algorithm are established to reduce the computational complexity of the C-index while accommodating larger sample sizes. The simulation study is conducted to examine the empirical performance of the proposed methodology. An analysis of SELECT data is carried out to further demonstrate the usefulness of the proposed joint models and model assessment criteria.

An Interpretable Bayesian Modeling Framework for Masked Competing Risk Survival Data with Cured Fractions

Shike Xu, University of Connecticut

Abstract: In addressing multiple risks in survival analysis, a subset of the population representing a "cured" fraction remains unaffected by all the causes under consideration. We introduce a novel competing risk modelling framework to incorporate an overall cure fraction, enhancing interpretability within this context and partially addressing the identifiability issue of using cause-specific cure fractions in a mixture model. Bayesian inference is conducted, utilizing Jeffreys' prior to further help address the identifiability issues of the cure rate mixture model. Collapsed Gibbs sampling is employed to improve mixing and convergence of Markov chain Monte Carlo (MCMC) sampling. To demonstrate the efficacy of our approach, we conduct a real data analysis using SEER data, comparing models with and without cure fractions. Additionally, we propose an overall Deviance Information Criterion (DIC) to evaluate model performance and serve as a benchmark for future comparison with other competing risk models that incorporate cure fractions.

Bayesian Inference of A Unified Estimand under Survival Models with Cure Fraction

Hongfei Li, Incyte Corporation

Abstract: In oncology research, cure models play a pivotal role in analyzing time-to-event data, particularly for diseases where a significant fraction of patients are cured. Among these, the Mixture Cure Model (MCM) and Promotion Time Cure Model (PTCM) are prominently utilized. We propose a unified estimand that describes the unconditional treatment effect for comparing treatment and control groups. This estimand focuses on whether the treatment extends survival for patients and connects with the conditional treatment effect parameters under different cure models. Moreover, this session will cover the Bayesian inference based on the proposed estimand, including Bayesian hypothesis testing and Bayesian model comparisons. Simulation studies demonstrate that, regardless of whether the model is correctly specified, the inference of the unified estimand yields desirable empirical performance. An analysis of the ECOG's melanoma dataset E1684 using the proposed estimand under different cure models is presented to illustrate the applicability of the proposed methodologies in real-world clinical research settings.

Advancements in Modern Inference for Correlated and Dependent Data (IS-61)

Chair: Jian Zou

Proposer: Jian Zou, Worcester Polytechnic Institute

Room: McHugh 305

Presenters: Adam Sales; Zheyang Wu; Yao Zheng; Jian Zou

Geepers: Principal Stratification Using Principal Scores and Stacked Estimating Equations

Adam Sales, Worcester Polytechnic Institute

Abstract: Principal stratification is a framework for making sense of causal effects conditioned on variables that themselves may have been affected by treatment. For instance, one component of an educational computer application is the availability of "bottom-out" hints that provide the answer. In evaluating a recent experimental evaluation against alternative programs without bottom-out hints, researchers may be interested in estimating separate average treatment effects for students who, if given the opportunity, would request bottom-out hints frequently, and for students who would not. Most principal stratification estimators rely on strong structural or modeling assumptions, and many require advanced statistical training to fit and check. In this paper, we introduce a new M-estimation principal effect estimator for one-way noncompliance based on a binary indicator. Estimates may be computed using conventional regressions (though the standard errors require a specialized sandwich formula) and do not rely on distributional assumptions. We present a simulation study that shows that the novel method is more robust than popular alternatives and illustrate the method in an analysis of data on bottom-out hint requests.

Asymptotic Correlation Robustness of Supremum-Based P-Value Combination Tests

Zheyang Wu, Worcester Polytechnic Institute

Abstract: A global hypothesis testing procedure is said to be asymptotically correlation robust if data correlation has a negligible influence on its type I error control at high significance levels. This property makes the testing procedure suitable for analyzing large datasets with complex correlations, such as whole genome sequencing studies (WGS). Such data analyses often require high significance levels, where correlation-robust tests can simplify computation while maintaining accurate type I error control through the independence approximation. A family of summation-based p-value combination tests using transformations of heavy-tailed distributions, such as the Cauchy Combination Test (CCT), have been shown to be asymptotically correlation robust. In this study, we prove that a broad family of supremum-based p-value combination tests also share this property. These include classic minP, Simes, Higher Criticism (HC), and phi-divergence tests. Additionally, we examine the non-asymptotic attributes of these tests across various linkage disequilibria among SNPs in WGS.

An Interpretable and Efficient Infinite-Order Vector Autoregressive Model for High-Dimensional Time Series

Yao Zheng, University of Connecticut

Abstract: As a special infinite-order vector autoregressive (VAR) model, the vector autoregressive moving average (VARMA) model can capture much richer temporal patterns than the widely used finite-order VAR model. However, its practicality has long been hindered by its non-identifiability, computational intractability, and difficulty of interpretation, especially for high-dimensional time series. This paper proposes a novel sparse infinite-order VAR model for high-dimensional time series, which avoids all above drawbacks while inheriting essential temporal patterns of the VARMA model. As another attractive feature, the temporal and cross-sectional structures of the VARMA-type dynamics captured by this model can be interpreted separately, since they are characterized by different sets of parameters. This separation naturally motivates the sparsity assumption on the parameters determining the cross-sectional dependence. As a result, greater statistical efficiency and interpretability can be achieved with little loss of temporal information. We introduce two ℓ_1-regularized estimation methods for the proposed model, which can be efficiently implemented via block coordinate descent algorithms, and derive the corresponding nonasymptotic error bounds. A consistent model order selection method based on the Bayesian information criteria is also developed. The merit of the proposed approach is supported by simulation studies and a real-world macroeconomic data analysis.

Boston-Pupa: A Bayesian Online Spatio-Temporal Outbreak Detection Framework with

Prior Updating and P-Value Adaptation

Jian Zou, Worcester Polytechnic Institute

Abstract: Early online outbreak detection for an epidemic is vital for disease-control authorities to make policies for the protection of public health and normal socioeconomic functions. Modern public health streaming surveillance data are often collected from multiple data sources, exhibiting spatio-temporal interdependence and imbalance issues. To address those issues, we propose a Bayesian online spatio-temporal outbreak detection with prior updating and p-value adaptation (BOSTON-PUPA) procedure. Using sequential p-value combinations, this iterative procedure involves the generalized Poisson distribution (GPD) model and supports synchronous surveillance over multiple locations, with a controlled false detection rate as well as high sensitivity against outbreaks in a wide range of signal-to-noise ratios. In the simulation study, we employed and compared several popular combined p-value methods in the BOSTON-PUPA procedure based on sensitivity, specificity, false detection rate and delay before making recommendations. We illustrated our method by detecting the outbreaks in the real COVID-19 daily case count data in Massachusetts counties in 2020.

Utility and Limitation of Re-Randomization Methods in Clinical Trials Research (IS-68)

Chair: Alex Sverdlov

Proposer: Yeh-Fong Chen, Food and Drug Administration

Room: McHugh 307

Presenters: Ke Zhu; Alex Sverdlov; Qing Liu

Analytical Approach to The Inversion of Fisher Randomization Tests

Ke Zhu, Duke University and North Carolina State University

Abstract: The Fisher randomization test (FRT) is advocated by many scholars because it produces finite-sample exact p-values for any test statistic and can be easily adapted to any experimental design. By inverting FRTs, we can construct the randomization-based confidence interval (RBCI). However, there are two main criticisms of randomization-based inference with RBCI. Firstly, when the test statistic does not satisfy certain monotonic conditions, the RBCI generated by existing numerical inversion approaches is questionable (Luo et al., 2021). Secondly, the exact coverage of the RBCI also requires the restricted condition of constant treatment effects. Wu and Ding (2021) proved that FRT using a studentized statistic is asymptotically valid for testing weak null hypotheses, and the corresponding RBCI asymptotically achieves nominal coverage probability without the condition of constant treatment effects. However, the studentized statistic, like many other commonly used test statistics, does not satisfy the monotonic conditions in Luo et al. (2021). This critical contradiction has hindered the practical application of RBCI. In this paper, we propose a general analytical approach to invert the FRT for test statistics that produce a non-monotonic p-value function, with the studentized t-statistic being an important special case. The RBCI generated by the proposed analytical approach is guaranteed to achieve the desired coverage probability and resolve the contradiction between Luo et al. (2021) and Wu and Ding (2021). Simulation results validate our findings and show that our method is computationally efficient.

Using Re-Randomization Rests to Address Disruptions in Clinical Trials

Oleksandr Sverdlov, Novartis

Abstract: Re-randomization tests can provide nonparametric inference that is robust to violation of the assumptions usually made in clinical trials. The ICH E9 (R1) Addendum on estimands and sensitivity analyses provides a guideline for aligning the trial objectives with strategies to address disruptions in clinical trials. In this presentation, we discuss a potential for embedding randomization tests within the estimand framework to allow for inference following disruptions in clinical trials in a way that reflects recent literature.

On Immortal Time Bias in Clinical Research with Survival Data

Qing Liu, Quantitative and Regulatory Medical Science (QRMedSci), LLC

Abstract: Survival analysis usually relies on time-to-event data arising from a counting process for which

the information is based on the number of events. We are interested in medical research to compare the survival of patients receiving a new treatment with the survival of patients receiving a control (e.g., placebo or standard-of-care). The most often used methods of inference are derived from the Cox proportional hazards model. These event-based methods run into serious problems when patients receiving the new treatment have zero or very few events (e.g., 1 or 2 events). In an estimand framework, we propose to summarize efficacy by averaging the survival rates from the KM curve for the treated patients at specified time points via an Inverse Probability Weighted Average (IPVA) method. We develop a many-to-1 index time (i.e., the initiation time of the new treatment) matching procedure by which the start time of each treated patient is matched by natural history controls from sources external to or by concurrent controls from the clinical research. Thus, each treated patient has a set of matched controls from whom survival data as measured from the index time of the treated patient being matched are well defined. The survival rates at the same specified time points are obtained from KM curve of all matched controls from which the efficacy of the matched controls is measured by the IPVA method. As a control patient can provide a match for different treated patients, the matched control data following the many-to-1 index time matching process are statistically dependent. Therefore, inference of the efficacy for the treated patients is conditional on the matched controls. The many-to-1 index time matching procedure is essential in comparative analysis of a new treatment with natural history controls for robust comparison against immortal time bias, which a form of selection bias first recognized by Gail (1972) by which patients need to survive long enough to receive a treatment of interest whereas there is no wait time for patients receiving a control. Immortal time bias is also present in randomized controlled trials for settings where patients are randomized to a treatment group and a control group. However, for patients randomized to the treatment group, the treatment of interest for individual patients is not available at the time of randomization, rather each patient has to go through a selection process of meeting certain outcome dependent criteria as an integral part of the treatment of interest. The presence of this outcome dependent selection process has the duo consequences of elimination of patients who do not meet the required criteria from the intent-to-treat population and enriching patients who receive the treatment of interest with favorable prognosis for treatment outcomes. Examples include but are not limited to scenarios where the treatment of interest 1) is not immediately available at the time of randomization (e.g., CART-T for cancer), 2) is contingent upon patient being successfully complete a pre-treatment procedure (e.g., surgery) during an initial run-in period, 3) requires a predicative biomarker of individual patients reaches a certain predetermined level (e.g., CD4 counts in HIV treatment), and 4) is individualized according to efficacy and safety comes for patients have successfully complete an initial treatment (e.g., individualized maintenance therapy following the results of an induction therapy in cancer). The proposed methodology is applicable to other many clinical trial settings with survival data where a current event (e.g., disease progression) needs to be confirmed by future data. This violation of the counting process has so far been ignored and the validity of statistical inference with existing statistical methods is largely unknown. The proposed methodology also addresses issues in certain applications where censoring is informative and potentially differential between the treatment and control groups.

Student Paper 8 (S8)

Chair: TBD

Organizer: Neil Spencer

Room: McHugh 306

Lool: Flexible & Robust Estimator of Heterogeneous Treatment Effects

Duy Minh Pham, Worcester Polytechnic Institute

Abstract: This thesis presents a two-stage approach – dubbed "Leave-One-Out Learner" (LOOL) – for estimating heterogeneous treatment effects – with the conditional average treatment effect (CATE) as the estimand of interest, as an extension of the "Leave-One-Out Potential Outcomes" (LOOP) Estimator proposed by Wu and Gagnon-Bartsch (2018). Researchers can first obtain unbiased estimates of the individual treatment effects (ITE) from the LOOP estimator and can then estimate the CATE by regressing on these estimated ITE. This estimator is robust as it guarantees unbiased estimates of the ITE in the first stage. It is also flexible since researchers can use any off-the-shelf machine learning methods for either stage without additional modification. More-

over, researchers can utilize parametric approaches like simple least-squares for drawing inferences on the impacts of each covariate or leverage more flexible non-parametric ones for more accurate predictions of the effects. We first compared this estimator to the meta-learner algorithms proposed by Kunzel et al. (2019) in initial simulation studies, where its performance is competitive. We then implemented and tested this method on data from a study examining the effectiveness of three educational platforms in teaching middle school algebra. The applied portion of this thesis is adapted from a paper in submission at the Seventeenth International Conference on Educational Data Mining (EDM 2024) by Duy Pham (the thesis's author), Kirk Vanacore, Adam Sales (the thesis's advisor), and Johann Gagnon-Bartsch.

Individualized Policy Evaluation and Learning under Clustered Network Interference

Yi Zhang, Harvard University

Co-authors: Kosuke Imai

Abstract: While there now exists a large literature on policy evaluation and learning, much of prior work assumes that the treatment assignment of one unit does not affect the outcome of another unit. Unfortunately, ignoring interference may lead to biased policy evaluation and ineffective learned policies. For example, treating influential individuals who have many friends can generate positive spillover effects, thereby improving the overall performance of an individualized treatment rule (ITR). We consider the problem of evaluating and learning an optimal ITR under clustered network interference (also known as partial interference) where clusters of units are sampled from a population and units may influence one another within each cluster. Unlike previous methods that impose strong restrictions on spillover effects, the proposed methodology only assumes a semiparametric structural model where each unit's outcome is an additive function of individual treatments within the cluster. Under this model, we propose an estimator that can be used to evaluate the empirical performance of an ITR. We show that this estimator is substantially more efficient than the standard inverse probability weighting estimator, which does not impose any assumption about spillover effects. We derive the finite-sample regret bound for a learned ITR, showing that the use of our efficient evaluation estimator leads to the improved performance of learned policies. Finally, we conduct simulation and empirical studies to illustrate the advantages of the proposed methodology.

On GEE for Mean-Variance-Correlation Models: Variance Estimation and Model Selection

Zhenyu Xu, University of Connecticut

Co-authors: Jason P. Fine, Wenling Song, Jun Yan

Abstract: Generalized estimating equations (GEE) are of great importance in analyzing clustered data without full specification of multivariate distributions. A recent approach jointly models the mean, variance, and correlation coefficients of clustered data through three sets of regressions (Luo and Pan, 2022). We observe that these estimating equations, however, are a special case of those of Yan and Fine (2004) which further allows the variance to depend on the mean through a variance function. The proposed variance estimators may be incorrect for the variance and correlation parameters because of a subtle dependence induced by the nested structure of the estimating equations. We characterize model settings where their variance estimation is invalid and show the variance estimators in Yan and Fine (2004) correctly account for such dependence. In addition, we introduce a novel model selection criterion that enables the simultaneous selection of the mean-scale-correlation model. The sandwich variance estimator and the proposed model selection criterion are tested by several simulation studies and real data analysis, which validate its effectiveness in variance estimation and model selection. Our work also extends the R package geepack with the flexibility to apply different working covariance matrices for the variance and correlation structures.

Efficient Inference on High-Dimensional Linear Models with Missing Outcomes

Yikun Zhang, University of Washington

Co-authors: Alexander Giessing, Yen-Chi Chen

Abstract: This paper is concerned with inference on the regression function of a high-dimensional linear model when outcomes are missing at random. We propose an estimator which combines a Lasso pilot estimate of the regression function with a bias correction term based on the weighted residuals of the Lasso regression. The weights depend on estimates of the missingness probabilities (propensity scores) and solve a convex optimization program that trades off bias and variance optimally. Provided that the propensity scores can be pointwise consistently estimated at in-sample data points, our proposed estimator for the regression function is asymptotically normal and semi-parametrically efficient among all asymptotically

linear estimators. Furthermore, the proposed estimator keeps its asymptotic properties even if the propensity scores are estimated by modern machine learning techniques. We validate the finite-sample performance of the proposed estimator through comparative simulation studies and the real-world problem of inferring the stellar masses of galaxies in the Sloan Digital Sky Survey.

New Pilot Study Design on Functional Data Analysis

Ping-Han Huang, Arizona State University

Abstract: Sparse functional measurements in practice are often contaminated with errors and collected at irregularly sampled time points. Extant work in response to this issue focuses on refining model estimation and the recovery of underlying trajectory, which heavily relies on the quality of a pilot study for the design of experiments. Our work aims to formulate a good pilot-study design to facilitate identifying optimal designs for future data collection and making statistical inference from pilot studies. We propose a design that helps find the best time points in the domain, minimizing the prediction errors for future subjects and the mean integrated squared errors for current subjects. A search algorithm is also developed to generate such pilot study and bring convenience for augmenting an existing design with additional subjects. Our design is shown the best by simulation for offering the smallest median and interquartile range of the bi-objective criterion compared to other designs. It is thus particularly useful for generating informative pilot data set that gives promising preliminary results and attracts further investments in developing full-scale studies.