**Abstract**

**Session 24CHIKT1: Opening Ceremony and Keynote Lecture 1&2**

**Robust Knockoffs Inference with Coupling**

♦*Yingying Fan*

University of Southern California

We investigate the robustness of the model-X knockoffs framework with respect to the misspecified or estimated feature distribution. We achieve such a goal by theoretically studying the feature selection performance of a practically implemented knockoffs algorithm, which we name as the approximate knockoffs (ARK) procedure, under the measures of the false discovery rate (FDR) and family wise error rate (FWER). The approximate knockoffs procedure differs from the model-X knockoffs procedure only in that the former uses the misspecified or estimated feature distribution. A key technique in our theoretical analyses is to couple the approximate knockoffs procedure with the model-X knockoffs procedure so that random variables in these two procedures can be close in realizations. We prove that if such coupled model-X knockoffs procedure exists, the approximate knockoffs procedure can achieve the asymptotic FDR or FWER control at the target level. We showcase three specific constructions of such coupled model-X knockoff variables, verifying their existence and justifying the robustness of the model-X knockoffs framework. This is a joint work with Jinchi Lv and Lan Gao.

**Deep regression learning with optimal loss function**

♦*Huanzhen Lin*

Southwestern University of Finance and Economics

In this paper, we develop a novel efficient and robust nonparametric regression estimator under a framework of a feedforward neural network (FNN). There are several interesting characteristics for the proposed estimator. First, the loss function is built upon an estimated maximum likelihood function, which integrates the information from observed data as well as the information from the data structure. Consequently, the resulting estimator has desirable optimal properties, such as efficiency. Second, different from the traditional maximum likelihood estimation (MLE), the proposed method avoids the specification of the distribution and thus is flexible to any kind of distribution, such as heavy tails and multimodal or heterogeneous distributions. Third, the proposed loss function relies on probabilities rather than direct observations as in least square loss, hence contributing to the robustness of the proposed estimator. Finally, the proposed loss function involves a nonparametric regression function only. This enables the direct application of the existing packages, simplifying the computational and programming requirements. We establish the large sample property of the proposed estimator in terms of its excess risk and minimax near-optimal rate. The theoretical results demonstrate that the proposed estimator is equivalent to the true MLE where the density function is known. Our simulation studies show that the proposed estimator outperforms the existing methods in terms of prediction accuracy, efficiency and robustness. Particularly, it is comparable to the true MLE and even gets better as the sample size increases. This implies that the adaptive and data-driven loss function from the estimated density may offer an additional avenue for capturing valuable information. We further apply the proposed method to four real data examples, resulting in significantly reduced out-of-sample prediction errors compared to existing methods.

**Session 24CHI6: Advanced Statistical Methods for Medical Data Analysis**

**Modeling and estimating a threshold effect: An application to improving cardiac surgery practices**

♦*Guangyu Yang*1*, Baqun Zhang*2*, Jonathan Haft*3*, Robert Hawkins*3*, David Sturmer*3*, Donald Likosky*3*, Min Zhang*3

1Renmin University of China，2Shanghai University of Finance and Economics， 3University of Michigan

Estimating thresholds when a threshold effect exists has important applications in biomedical research. However, mod- els/methods commonly used in the biomedical literature may lead to a biased estimate. For patients undergoing coronary artery bypass grafting (CABG), it is thought that exposure to low oxygen delivery (DO2) contributes to an increased risk of avoidable acute kidney injury. This research is motivated by estimating the threshold of nadir DO2 for CABG patients to help develop an evidence-based guideline for improving cardiac surgery practices. We review several models (sudden-jump model, broken-stick model, and the constrained broken-stick model) that can be adopted to estimate the threshold and discuss modeling assumptions, scientific plausibility, and implications in estimating the threshold. Under each model, various estimation methods are studied and compared. In particular, under a constrained broken-stick model, a modified two-step Newton–Raphson algorithm is introduced. Through comprehensive simulation studies and an application to data on CABG patients from the University of Michigan, we show that the constrained broken-stick model is flexible, more robust, and able to incorporate scientific knowledge to improve efficiency. The two-step Newton–Raphson algorithm has good computational performances relative to existing methods.

**Advances and Opportunities in Radiomics for Clinical Cancer Diagnosis and Treatment**

♦*Mengjie Fang*1 *and Di Dong*2

1Beihang University， 2Chinese Academy of Sciences

In recent years, an increasing body of research has indicated that Radiomics, which leverages medical imaging and artificial intelligence, can effectively address a range of clinical tasks associated with cancer, including screening, diagnosis, treatment, and prognosis. Gastric cancer and nasopharyngeal cancer, prevalent in China, significantly endanger public health. Herein, we will first outline the foundational techniques and workflows of Radiomics in the context of clinical cancer diagnosis and treatment. Then, we will specifically focus on gastric and nasopharyngeal cancers to illustrate the potential clinical impact of Radiomics. Moreover, we will discuss the challenges in methodological development and practical implementation, and explore future directions for potential breakthroughs in this field.

**Statistical Consolidation of Multiple Noisy Crowd Labels with Confidence**

♦*Qianhan Zeng*1*, Rui Pan*2*, Hansheng Wang*1

1Peking University， 2Central University of Finance and Economics

We propose a statistical model to predict the labels for the remaining unlabeled dataset. The model is constructed under a hybrid human-machine annotation scheme to save human efforts. First, a pilot but consistent estimator of the model parameters can be obtained based on the pilot sample and its crowd labels. Then, we rigorously prove the consistency and asymptotic normality of the proposed estimators. Third, once the pilot estimators are computed, they are then used to predict the labels for the remaining unlabeled data. Finally, comprehensive numerical studies are presented, including real-world labeling case studies, to demonstrate the effectiveness of our proposed framework.

**Bayesian Scalar-on-Image Regression with the Spatially Varying Neural Network Prior**

♦*Ben Wu*1*, Keru Wu*2*, Hansheng Wang*3

1Renmin University of China， 2Duke University， 3University of Michigan

Deep neural networks (DNN) have been adopted in the scalar-on-image regression which predicts the outcome variable using image predictors. However, training DNN often requires a large sample size to achieve a good prediction accuracy and the model fitting results can be difficult to interpret. In this work, we propose a noval Bayesian non-linear scalar-on-image regression framework with a spatially varying neural network SV-NN prior. The SV-NN is constructed using a single hidden layer neural network with its weights generated by the soft-thresholded Gaussian process. Our framework is able to select interpretable image regions and to achieve high prediction accuracy with limited training samples. The SV-NN provides large prior support for the imaging effect function, enabling efficient posterior inference on image region selection and automatically determining the network structures. We establish the posterior consistency of model parameters and selection consistency of image regions when the number of voxels/pixels grows much faster than the sample size. We develop an efficient posterior computation algorithm based on stochastic gradient Langevin dynamics (SGLD). We compared our methods with state-of-the-art deep learning methods via analyses of multiple real data sets including the task fMRI data in the Adolescent Brain Cognitive Development (ABCD) study.

**Session 24CHI28: Designs for Computer Experiments**

**Construction of Orthogonal-MaxPro Latin Hypercube Designs**

*Yaping Wang*1*, Sixu Liu*1*,* ♦*Qian Xiao*2

1East China Normal University， 2Shanghai Jiao Tong University

Orthogonal Latin hypercube designs (LHDs) and maximum projection (MaxPro) LHDs are widely used in computer experiments. They are efficient for estimating the trend part and the Gaussian process part of the universal Kriging (i.e. the Gaussian process) model, respectively, especially when only some of the factors are active. Yet, the orthogonality and the MaxPro criteria often do not agree with each other. In this work, we propose a new class of optimal designs, called orthogonal-MaxPro LHDs, optimizing a well-defined multi-objective criterion combining the correlation and the MaxPro metrics. An efficient parallel algorithm via level permutations and expansions is developed, whose efficiency is guaranteed by theories. Numerical results are presented to show that the construction is fast and the obtained designs are attractive, especially for large computer experiments.

**Selecting strong orthogonal arrays by linear allowable level permutations**

♦*Guanzhou Chen*1 *and Boxin Tang*2

1Nankai University,2Simon Fraser University.

Space-filling designs are widely used in physical and computer experiments when the model between the response and input factors is uncertain. Recently, Chen and Tang (2022, Ann. Statist. 50, 2925-2949) justified the use of strong orthogonal arrays (SOAs) under a broad class of space-filling criteria. However, when allowable level permutations are applied to an SOA, a class of SOAs can be obtained with different geometrical structures and it is not clear which one should be selected for practical use. In this paper, we address this issue by considering a representative subset of allowable level permutations, called linear allowable level permutations. These special level permutations offer theoretical convenience in classifying various geometrically non-isomorphic SOAs. Based on these results, construction methods are provided to obtain SOAs that are more space-filling than those in the literature.

**Grouped orthogonal arrays for computer experiments**

♦*Wenlong Li*1*,* *Jian-Feng*2*,* Peter Chien3

1Beijing Jiaotong University, 2Nankai University, 3University of Wisconsin-Madison

We propose a method for constructing a new type of space-filling design, called grouped orthogonal array, to accommodate natural input grouping in computer experiments. Such a design achieves space-filling properties in all inputs and possesses stronger space-filling properties between inputs in the same groups than between inputs from different groups. Using combinatorial orthogonality as the guiding space-filling criterion, our method generates a class of grouped orthogonal arrays that are special strength-two orthogonal array with columns partitioned into groups of strength-three orthogonal arrays. The proposed methods are easy to implement and can accommodate a large number of factors. Examples are presented to illustrate the methods. Numerical comparison is provided to present the usefulness of the grouping approach. Simulations are given to illustrate the effectiveness of the proposed designs.

**Optimal subsampling for Linear Mixed Models**

*Jiaqing Zhu*1, *LinWang*2, ♦*Fasheng Sun*3

1Northeast Normal University, 2Purdue University, 3Northeast Normal University

Hierarchical data analysis is crucial in various fields for making discoveries. The linear mixed model is often used for training hierarchical data, but its parameter estimation is computationally expensive, especially with big data. Subsampling techniques have been developed to address this challenge. However, most existing subsampling methods assume homogeneous data and do not consider the possible heterogeneity in hierarchical data. To address this limitation, we develop a new approach called group-orthogonal subsampling (GOSS) for selecting informative subsets of hierarchical data that may exhibit heterogeneity. GOSS selects subdata with balanced data size among groups and combinatorial orthogonality within each group, resulting in subdata that are D- and A-optimal for building linear mixed models. Estimators of parameters trained on GOSS subdata are consistent and asymptotically normal. GOSS is shown to be numerically appealing via simulations and a real data application.

**Session 24CHI34: High Dimensional Data Analysis**

**Semiparametric Estimation of Non-Ignorable Missingness with Refreshment Sample**

♦*Jing Wang*1 *and Elena Graetz*2

This research addresses the challenges of missing data in longitudinal studies, emphasizing the impact on unbiased estimation and statistical inference due to informative missing. Traditional methods assume Missing Completely at Random (MCAR) or Missing at Random (MAR), but we focus on non-ignorable missingness using refreshment samples in two-wave panel data. The study proposes an additive attrition model with B-spline estimation for missingness components based on identification equations by Hirano et al. (2001). We demonstrate the convergence of the objective function to its probability limit and the consistency of the spline estimator. Large sample behavior is compared with linear and partially linear attrition models. Application to the Netherlands Mobility Panel and Understanding Society dataset explores the relationship between missingness and income. The findings highlight the necessity of understanding attrition processes for selecting appropriate statistical tools in longitudinal studies.

**TBD**

♦*Xu Liu*1

TBD

**A Unified Approach to Variable Selection for Partially Linear Models**

*Youhan Lu*1, *Yushen Dong*1, ♦*Juan Hu*2 ， *Yichao Wu*1.

1University of Illinois Chicago， 2DePaul University.

We focus on the general partially linear model without any structure assumption on the nonparametric component. For such a model with both linear and nonlinear predictors being multivariate, we propose a new variable selection method. Our new method is a unified approach in the sense that it can select both linear and nonlinear predictors simultaneously by solving a single optimization problem. We prove that the proposed method achieves consistency. Both simulation examples and a real data example are used to demonstrate the new method’s competitive finite-sample performance.

**Paralinear distance and its algorithm for hierarchical clustering of high-dimensional discrete variables**

*Shuai Wang1, Lizhu Hao1,* ♦*Xiaofei Wang1, Jianhua Guo2*

1KLAS and School of Mathematics and Statistics, Northeast Normal University 2School of Mathematics and Statistics, Beijing Technology and Business University

Variable clustering is an important tool for mining association rules and explaining the latent mechanisms responsible for generating data. In this work, we aim to study the hierarchical variable clustering algorithm based on the paralinear distance between discrete variables. Firstly, we study the paralinear distance with the multinomial distribution, and point out that any

distance with additivity on the graphical tree model has a unique form on the paralinear distance. And then, we suggest a novel hierarchical clustering algorithm, which can determine the local relationships of observed variables as sibling groups and singletons in each level, where the hierarchical structures are indicated between the levels. Furthermore, we show the probably approximately correct (PAC) property of the algorithm, and find out that its sample complexity is sensitive to the diameter of the tree. Finally, by using GPU computation, we demonstrate our discoveries and the applications of our learning algorithms through large-scale experiments on both synthetic and real-world data. Extensive empirical results show that the proposed method is efficient for discovering local structures and latent information.

**Session 24CHI41: Lifetime Data Analysis**

**An Instrumental Variable Approach for the Doubly Robust Estimation of Average Treatment Effects in Time-to-Event Data with Unmeasured Confounding**

♦*Chung-Chou Chang*1*， Runjia Li*1*.*

1University of Pittsburgh

Observational studies are widely used for identifying causal treatment effects, yet often suffer from biased results due to the violation of the unconfoundedness assumption. Instrumental variable methods have been developed for estimating causal effects in the presence of unmeasured confounding. However, existing approaches fall short in estimating average treatment effects (ATE) for time-to-event outcomes in observational data, and often being constrained to specific survival model and lacking in desired statistical properties. In this study, we introduce a novel instrumental variable estimator of ATE in time-to-event outcomes, based on the cumulative incidence functions, accommodating scenarios with or without competing risks. Derived from efficient influence function, our estimator possesses double robustness and asymptotic efficiency, which were theoretically proved and demonstrated via simulations. Our method allows for the incorporation of various models for outcome, treatment and censoring, including machine learning and ensemble learning algorithms.

**Semiparametric additive hazards models for recurrent gap times in the presence of a terminal event: an application to a recurrent thyroid cancer study**

*Yong Huang*1, *Deng-Huang Su*2, ♦*Shu-Hui Chang*1.

1National Taiwan University,2Ming-Ren Clinic.

Recurrent event data in presence of a terminal event, such as death, are frequently encountered in epidemiological and clinical studies. Gap time from therapeutic intervention to the next event, recurrent event or terminal event, is often the natural outcome of interest in practices. The additive hazards models for successive gap times are considered where the additive effects are applicable to composite and separate endpoints of recurrent and terminal events. Weighted estimation methods for covariate effects are developed under dependent censoring mechanism. In addition, formal tests for covariate effects independent of event types, gap episodes and time are developed. The asymptotic properties of proposed methods are studied. Simulation studies are also conducted to assess their finite-sample performance. Finally, our methods are applied to a thyroid cancer study.

**First-hitting-time Threshold Regression for Longitudinal Time-to-event Data**

♦*Mei-Ling Ting*1 *and George Whitmore*2

1University of Maryland College Park 2McGill University

Disease progression in a patient can be described mathematically as a stochastic process. The patient experiences a failure event when his/her disease progression first reaches a critical threshold level. This happening defines the failure event itself and the first hitting time (FHT) is the event time. First hitting time threshold regression (TR) models incorporate regression functions for parameters of the underlying stochastic process. These TR models do not require the proportional hazards assumption and represent a realistic alternative to the Cox model for capturing granular structure in a high-dimensional model. In addition to parametric models, FHT TR models have recently been extended to semiparametric applications. Using the technique of Markov decomposition, the FHT TR model can handle longitudinal time-to-event data. We demonstrate the application using the Osteoarthritis Initiative (OAI) data.

**Efficient auxiliary information synthesis for cure rate model**

*Jie Ding*1, ♦*Jialiang Li*2 *and Xiaoguang Wang*1.

1Dalian University of Technology， 2National University of Singapore.

We propose a new auxiliary information synthesis method to utilize subgroup survival information at multiple time points under the semi-parametric mixture cure rate model. After summarizing the auxiliary information via estimating equations, a control variate technique is adopted to reduce the variance efficiently, together with a test statistic to check the homogeneity assumption. Revision using penalization is further considered to adaptively accommodate potential population heterogeneity. Our methods can be adjusted when the uncertainty is not negligible. We establish asymptotic properties of our proposed estimators, and

demonstrate their practical performances through extensive simulations and an invasive breast cancer study.

**Session 24CHI53: New Nonparametric and Dimension Reduction Methods for Causal Inference and Clustering Analysis**

**An exhaustive selection of sufficient adjustment sets for causal inference**

♦*Wei Luo*1*, Fei Qin*2 *and Li-Xing Zhu*3*.*

1Zhejiang University 2HongKong Baptist University 3Beijing Normal University.

A subset of predictors that satisfies the ignorability assumption, called the sufficient adjustment set, is crucial for conducting reliable causal inference based on observational data. In this talk, we propose a general family of methods to detect all such sets for the first time in the literature, with no parametric assumptions on the outcome models and with flexible parametric and semiparametric assumptions on the predictors within the treatment groups; the latter induces desired sample-level accuracy. We show that the collection of sufficient adjustment sets can uniquely facilitate multiple types of studies in causal inference, including sharpening the estimation of average causal effect and recovering fundamental connections between the outcome and the treatment hidden in the dependence between predictors. These findings are illustrated by simulation studies and a real data example at the end.

**A Bayesian nonparametric approach for handling item and examinee heterogeneity in assessment data**

♦*Guanyu Hu*1

1The University of Texas Health Science Center at Houston

In this talk, I will introduce a novel nonparametric Bayesian item response theory model that estimates clusters at the question level, while simultaneously allowing for heterogeneity at the examinee level under each question cluster, characterized by a mixture of binomial distributions. The main contribution of this work is threefold. First, we present our new model and demonstrate that it is identifiable under a set of conditions. Second, we show that our model can correctly identify question-level clusters asymptotically, and the parameters of interest that measure the proficiency of examinees in solving certain questions can be estimated at a sqrt{n} rate (up to a log term). Third, we present a tractable sampling algorithm to obtain valid posterior samples from our proposed model. Compared to the existing methods, our model manages to reveal the multi-dimensionality of the examinees' proficiency level in handling different types of questions parsimoniously by imposing a nested clustering structure. The proposed model is evaluated via a series of simulations as well as apply it to an English proficiency assessment data set. This data analysis example nicely illustrates how our model can be used by test makers to distinguish different types of students and aid in the design of future tests.

**Inference of continuous treatment effects in large-scale observational data**

♦*Shujie Ma*1

1UC-Riverside

Recent advances in technology have created numerous large-scale datasets in observational studies, which provide unprecedented opportunities for evaluating the effectiveness of various treatments. Under the condition of unconfounded treatment assignment, most existing methods rely on a parametric or a non-parametric modeling method for estimating the propensity score or the outcome regression functions. The parametric approach lacks robustness as it suffers from the model misspecification problem. Conventional nonparametric estimation methods suffer from the curse of dimensionality when the dimension of confounders is large. In this talk, I will introduce a new method we have proposed for estimating and inferring continuous treatment effects in large-scale observational data. Our nuisance function is estimated by artificial neural networks. Our proposed method takes full advantage of the large sample size of large-scale data and provides effective protection against misspecification bias. I will also present the theoretical properties we have established for the proposed estimator and illustrate the method through simulation studies and real data applications.

**Kernel-based Decentralized Policy Evaluation for Reinforcement Learning**

*Jiamin Liu*1 ，♦*Heng Lian*2

1University of Science and Technology Beijing, 2City University of Hong Kong

We investigate the decentralized nonparametric policy evaluation problem within reinforcement learning, focusing on scenarios where multiple agents collaborate to learn the state-value function using sampled state transitions and privately observed rewards. Our approach centers on a regression-based multi-stage iteration technique employing infinite-dimensional gradient descent within a reproducing kernel Hilbert space (RKHS). To make computation and communication more feasible, we employ Nystrom approximation to project this space into a finite-dimensional one. We establish statistical error bounds to describe the convergence of value function estimation, marking the first instance of such analysis within a fully decentralized nonparametric framework.

**Session 24CHI57: Public Health Practices Under Data Science**

**Blockchain-Enabled Social Media Network for Mitigating Misinformation Spread**

♦*Rui Luo*1*, Vikram Krishnamurthy*2*， Erik Blasch*3*.*

1City University of Hong Kong 2Cornell University 3Air Force Research Laboratory

In this talk, we present a novel blockchain protocol, known as BE-SMN (Blockchain-Enabled Social Media Network), designed to address the rampant spread of misinformation in online platforms. Our approach leverages the information transmission-time distribution and models the transmission of misinformation as double-spend attacks on a blockchain framework. By integrating the misinformation distribution into the classic SIR (Susceptible, Infectious, or Recovered) model, we introduce a modified SIR model that replaces the traditional single rate parameter. This adaptation allows us to analyze the impact of misinformation propagation within a multi-community network. Through numerical simulations, we demonstrate the effectiveness of our proposed blockchain-enabled social media network in curbing the spread of misinformation. Specifically, we showcase how our network outperforms a baseline network by effectively flattening the curve of the infected population. This talk offers insights into combating the pervasive issue of misinformation in social media platforms. By leveraging blockchain technology, we aim to contribute to the development of robust and reliable mechanisms for safeguarding the integrity of information dissemination in online communities.

**Unlocking the potential of statistical engineering in precision healthcare**

♦*Li Tang*1

1St. Jude Children's Research Hospital

In the rapidly evolving field of precision healthcare, the integration of statistical engineering principles presents a transformative opportunity to enhance patient outcomes and optimize treatment strategies. This study explores the innovative application of statistical engineering techniques to address complex challenges within precision healthcare, aiming to refine diagnostic accuracy, tailor treatment interventions, and predict patient responses with unprecedented precision. By harnessing advanced analytics, machine learning algorithms, and data-driven decision-making processes, we demonstrate how statistical engineering can significantly contribute to the personalization of healthcare. Furthermore, this approach facilitates a deeper understanding of the underlying mechanisms of diseases, enabling healthcare professionals to make more informed decisions and offer treatments that are specifically designed to meet the unique needs of individual patients. The promising results of this study underscore the vital role of statistical engineering in unlocking the full potential of precision healthcare, suggesting a pathway toward more effective and personalized medical care.

**Survival Bandits**

♦*Yinghao Pan*1*, Eric Labe*2 *， Yingqi Zhao*3

1University of North Carolina at Charlotte， 2Duke University， 3Fred Hutchinson Cancer Center

We consider a contextual survival bandit setting, a variant of the classical multi-armed bandit problem in which the reward for each individual is a survival time subject to right censoring. First, we design a Thompson sampling algorithm that randomly allocates individuals to treatments in an adaptive manner based on Bayesian posterior distributions. Next, we propose a weighted M-estimator for constructing valid confidence regions using data collected from the Thompson sampling algorithm mentioned above. Asymptotic properties of the proposed weighted M-estimator are established by careful use of martingale theory.

**On mixed sampling under unidentified nonpaprametric models: Application to the transformation model**

*Chong Zhong*1, *Jin Yang*2, *Junshan Shen*3, ♦*Catherine Liu*1, *Zhaohai Li*4

1The Hong Kong Polytechnic University,

2Enice Kennedy Shriver National Institute of Child Health and Human Development, NIH, USA，3Capital University of Economics and Business, China，4Geoge Washington University, USA.

Poor mixing is a stumbling block to Bayesian prediction under nonparametric models with multiple unidentified infinite-dimensional parameters, incurring poor estimation of the posterior predictive distribution. Unlike existing strategies of modifying the samplers or processing the posterior, we address the poor mixing by establishing a conception of sufficiently informative priors, which guarantee the sampled posterior reflects the posterior variation correctly. Specifically, under the nonparametric transformation model, we provide a criterion of sufficient informativeness with a threshold that formulates the asymptotic posterior variance through the hyperparameters in nonparametric priors.

**Session 24CHI67: Recent Advances in Modeling Single-Cell Data**

**Identify Causal Gene Regulation Using Observational Single-Cell Data**

♦*Yang Ni*1*.*

1Texas A&M University

Using observational data to identify causal gene regulation is a challenging task. In this talk, we introduce a new method to discover new causal gene regulatory relationships based on observational single-cell data.

**Detecting Circadian Rhythmicity by Single-Cell Multiomics**

*Anthony Tong*1*, Jerome Menet*1*，* ♦*Yuchao Jiang* 1

1Texas A&M University

From bacteria to humans, most organisms showcase inherent 24-hour circadian rhythms. Best exemplified with the sleep-wake cycle, these rhythms are remarkably widespread, governing hormonal, metabolic, physiological, and behavioral oscillations and enabling most biological functions to perform optimally at the most appropriate time of the day. Circadian rhythms are generated by “molecular clocks” that drive the rhythmic expression of thousands of genes throughout the body. Recent advances in next-generation sequencing (NGS) technologies offer appealing platforms to identify circadian genes and elucidate the mechanisms that underlie their rhythmic expression. In a typical high-throughput experiment, bulk tissues from organisms such as mouse, fruit flies, or plants are collected at six to eight different circadian time points, followed by NGS such as RNA sequencing to generate time-series expression profiles for all genes across the genome. While significant efforts have yielded notable successes, recent advances in single-cell multiomics technologies have further enabled the exploration of circadian rhythmicity with a much finer resolution. This surge, however, necessitates the development of novel methodologies. In this presentation, I will discuss our experimental, computational, and statistical methods on detecting circadian rhythmicity using single-cell RNA + ATAC multiomics data.

**Regression Analysis for Single-cell RNA-seq Data**

♦*Fangda Song*1*, Kevin Y. Yip*2 *， Yingying Wei*1

1The Chinese University of Hong Kong 2Sanford Burnham Prebys Medical Discovery Institute

scRNA-seq studies that assay a large cohort of donors are emerging, which provides opportunities for us to understand how gene expression profiles of a cell type are affected by donors’ characteristics, such as age, gender and disease status. However, statistical methods developed to study the association between bulk gene expression data and a set of covariates are not applicable to scRNA-seq data, because the cell-type label of each cell is unknown. In addition, batch effects, variations in cell-type abundance between donors, and missing data due to dropout events all add challenges to detecting the association between scRNA-seq data and covariates. Here, we develop Regress-seq, a Bayesian hierarchical model that can simultaneously cluster cell types, correct batch effects and infer the covariates effects on cell-types-specific gene expression profiles. Moreover, we derive the conditions of the experimental designs under which the integrative analysis of multiple scRNA-seq studies is valid so that the cell type effects, batch effects and covariate effects can be separated. We envision Regress-seq to greatly facilitate the development of personalized medicine.

**Exploring Advanced Frontiers in Single-Cell Bioinformatics: Moving Beyond Differential Expression Analysis**

♦*James Cai*1

1Texas A&M University

Differential expression (DE) analysis, a cornerstone of gene function studies, often yields inconsistent results and detects genes with limited biological relevance. This challenge is particularly acute in single-cell transcriptomics, a rapidly growing field. Dr. James Cai will address the underlying limitations of DE and introduce data-driven methods to extract meaningful biological insights from single-cell data. These methods go beyond DE and include: (1) Gene regulatory network construction and comparison [1]—this approach reveals how genes interact and identifies key regulators in different cellular contexts; (2) Virtual gene knockout analysis [2]—this technique simulates gene inactivation to predict functional consequences and identify essential genes; and (3) Manifold learning analysis for cell-cell communication [3]—this method uncovers communication patterns between cells, providing insights into cellular interactions. By moving beyond DE, these approaches offer a deeper understanding of gene function and cellular processes in single-cell transcriptomics data.

**Session 24CHI70: Recent Advances in Network Data Analysis**

**Privacy-Protected Spatial Autoregressive Model**

♦*Danyang Huang*1*, Ziyi Kong*1*, Shuyuan Wu*2 *， Hangsheng Wang*3

1Renmin University of China， 2Shanghai University of Finance and Economics 3Peking University

Spatial autoregressive (SAR) models are important tools for studying network effects. However, with an increasing emphasis on data privacy, data providers often implement privacy protection measures. This makes classical SAR models inapplicable. In this work, we introduce a privacy-protected SAR model with noise-added response and covariates to meet privacy protection requirements. However, in this scenario, the traditional quasi maximum likelihood estimator (QMLE) becomes infeasible because the likelihood function cannot be formulated. To tackle this issue, we first consider an explicit expression for the likelihood function with noise-added response only. Unfortunately, the derivatives are biased due to the noise in covariates. Then, we develop techniques to correct the biases introduced by noise. Correspondingly, a Newton-Raphson type algorithm is proposed to obtain the estimator. This leads to a corrected likelihood estimator (CLE). To further enhance computational efficiency, we also introduce a corrected least squares estimator (CLS) based on the idea of bias correction. These two estimation methods ensure both data security and the attainment of statistically valid estimators. Theoretical analysis of both estimators has been carefully established. Statistical inference methods have also been carefully discussed. The finite sample performances of different methods are demonstrated through extensive simulations and the analysis of a real dataset.

**Testing degree heterogeneity in directed networks**

♦*Lu Pan*1*, Ting Yan*1 *， Qiuping Wang*2

1 Central China Normal University， 2Zhaoqing University

Wilks' theorems state that the likelihood ratio test statistic has an asymptotic chi-square distribution, which is one of the most familiar of results about maximum likelihood. In this paper, we study the degree heterogeneity testing method in the $p\_0$ model for directed networks.Let $\ell( \bs{\theta})$ be the log-likelihood function, $\widehat{\bs{\theta}}$ be the maximum likelihood estimator (MLE) of $\bs\theta$,and $\widehat{\bs{\theta}}^0$ be the restricted MLE for $\bs\theta$ under the null hypothesis $H\_{0}$. Consider growing dimensional null hypotheses $H\_{0}: \alpha\_{i}=\alpha\_{i}^{0}$ for $i=1,\ldots,r$ or $H\_{0}: \alpha\_{1}=\cdots=\alpha\_{r}$. Under the null hypothesis $H\_{0}$, the normalized log-likelihood ratio statistic, $[2\{\ell(\widehat{\bs\theta})-\ell(\widehat{\bs\theta}^{0})\}-r]/(2r)^{1/2}$, converges in distribution to a standard normal distribution, as $r\rightarrow \infty$. Under the homogenous null $H\_0: \alpha\_1=\cdots=\alpha\_r$ with afixed $r$, we prove that $2\{\ell(\widehat{\bs{\theta}}) - \ell(\widehat{\bs{\theta}}^0)\}$ converges in distribution to a Chi-square distribution with $r-1$ degrees of freedom as $n\rightarrow \infty$. It does not depend on the nuisance parameters. Our results can be interpreted as a high-dimensional extension of the Wilks theorem. For the null $H\_0: \alpha\_{i}=\alpha\_{i}^{0}$ with a fixed $r$, we discover a different phenomenon that the distribution of the log-likelihood ratio statistic does not converge to a Chi-square distribution. Numerical simulations illustrate our theoretical results, and we apply them to the Kapferer tailor shop dataset.

**Network-Guided Feature Selection and Downstream Applications**

♦*Wanjie Wang*1*， Tao Shen*1

1National University of Singapore

Nowadays, it is frequently seen that the data set often contains network information and high-dimensional covariates. Studies have shown that the covariates will help uncover the network structure. The opposite direction should also work, that the network information helps to denoise the covariates and improve the statistical inference. In this talk, I will present a covariate selection method based on the spectral information of the adjacency matrix. By the eigenvectors, we design a testing statistic of the covariates and select them by Higher-Criticism statistic. We prove the optimality of this method and its effect in the regression and clustering problems. We finally apply it to a sina dataset.

**A two-way heterogeneity model for dynamic networks**

♦*Binyan Jiang*1*, Chenlei Leng*2 *, Ting Yan*3 *, Qiwei Yao*4 *，Xinyang Yu*1

1 The Hong Kong Polytechnic University， 2 University of Warwick， 3 Central China Normal University， 4London School of Economics

Dynamic network data analysis requires joint modelling individual snapshots and time dynamics. This paper proposes a new two-way heterogeneity model towards this goal. The new model equips each node of the network with two heterogeneity parameters, one to characterize the propensity of forming ties with other nodes and the other to differentiate the tendency of retaining existing ties over time. Though the negative log-likelihood function is non-convex, it is locally convex in a neighbourhood of the true value of the parameter vector. By using a novel method of moments estimator as the initial value, the consistent local maximum likelihood estimator (MLE) can be obtained by a gradient descent algorithm. To establish the upper bound for the estimation error of the MLE, we derive a new uniform deviation bound, which is of independent interest. The usefulness of the model and the associated theory are further supported by extensive simulation and the analysis of some real network data sets.

**Session 24CHI72: Recent Advances in Statistical Genetics and Genomics**

**A knockoff framework for the identification of putative causal variants in genome-wide association studies with trio design**

♦*Yi Yang*1*, Chen Wang*2*, Linxi Liu*3*, Joseph Buxbaum*4 *，* *Zihuai He*5*.*

1City University of Hong Kong, 2Columbia University, 3University of Pittsburgh, 4Icahn School of Medicine at Mount Sinai and 5Stanford University

Family-based designs can eliminate confounding due to population substructure and can distinguish direct from indirect genetic effects, but these designs are underpowered due to limited sample sizes. Here, we propose KnockoffTrio, a statistical method to identify putative causal genetic variants for father-mother-child trio design built upon a recently developed knockoff framework in statistics. KnockoffTrio controls the false discovery rate (FDR) in the presence of arbitrary correlations among tests and is more powerful than the conventional methods that control the family-wise error rate via Bonferroni correction. Furthermore, KnockoffTrio is not restricted to family-based association tests and can be used in conjunction with more powerful, potentially nonlinear models to improve the power of standard family-based tests. We show, using empirical simulations, that KnockoffTrio can prioritize causal variants over associations due to linkage disequilibrium and can provide protection against confounding due to population stratification. In applications to 14,200 trios from three study cohorts for autism spectrum disorders (ASDs), including AGP, SPARK, and SSC, we show that KnockoffTrio can identify multiple significant associations that are missed by conventional tests applied to the same data.

**Improving cell type deconvolution accuracy using personalized reference profiles**

♦*Hao Feng*1

1 Case Western Reserve University

Real-world clinical samples are often admixtures of signal mosaics from multiple pure cell types. Using computational tools, bulk transcriptomics can be deconvoluted to solve for the abundance of constituent cell types. However, existing deconvolution methods are conditioned on the assumption that the whole study population is served by a single reference panel, which ignores person-to-person heterogeneity. Here we present imply, a novel algorithm to deconvolute cell type proportions using personalized reference panels. imply can borrow information across repeatedly measured samples for each subject, and obtain precise cell type proportion estimations. Simulation studies demonstrate reduced bias in cell type abundance estimation compared with existing methods. Real data analyses on large longitudinal consortia show more realistic deconvolution results that align with biological facts. Our results suggest that disparities in cell type proportions are associated with several disease phenotypes in type 1 diabetes and Parkin-son’s disease. Our proposed tool imply is available through the R/Bioconductor package ISLET at <https://bioconductor.org/packages/ISLET/>.

**Continually adapting pre-trained language model to universal annotation of single-cell RNA-seq data.**

*Hui Wan*1*, Musu Yuan*1*, Yiwei Fu*1 *，* ♦*Minghua Deng*

1Peking University

Cell-type annotation of single-cell RNA-sequencing (scRNA-seq) data is a hallmark of biomedical research and clinical application. Current annotation tools usually assume the simultaneous acquisition of well-annotated data, but without the ability to expand knowledge from new data. Yet, such tools are inconsistent with the continuous emergence of scRNA-seq data, calling for a continuous cell-type annotation model. In addition, by their powerful ability of information integration and model interpretability, transformer-based pre-trained language models have led to breakthroughs in single-cell biology research. Therefore, the systematic combining of continual learning and pre-trained language models for cell-type annotation tasks is inevitable.

We herein propose a universal cell-type annotation tool, called CANAL, that continuously fine-tunes a pre-trained language model trained on a large amount of unlabeled scRNA-seq data, as new well-labeled data emerges. CANAL essentially alleviates the dilemma of catastrophic forgetting, both in terms of model inputs and outputs. For model inputs, we introduce an experience replay schema that repeatedly reviews previous vital examples in current training stages. This is achieved through a dynamic example bank with a fixed buffer size. The example bank is class-balanced and proficient in retaining cell-type-specific information, particularly facilitating the consolidation of patterns associated with rare cell types. For model outputs, we utilize representation knowledge distillation to regularize the divergence between previous and current models, resulting in the preservation of knowledge learned from past training stages. Moreover, our universal annotation framework considers the inclusion of new cell types throughout the fine-tuning and testing stages. We can continuously expand the cell-type annotation library by absorbing new cell types from newly arrived, well-annotated training datasets, as well as automatically identify novel cells in unlabeled datasets. Comprehensive experiments with data streams under various biological scenarios demonstrate the versatility and high model interpretability of CANAL.

**Inferring cell division tree from single-cell lineage barcode and single-cell transcriptomic data.**

♦Jiajun Zhang1， Xiaochen Yu1.

1Sun Yat-sen University.

The process of cell division and differentiation is crucial for the growth and development of multicellular organisms. Integrating single-cell RNA sequencing with CRISPR-Cas9 barcode editing has enabled lineage tracing and cell division tree reconstruction. However, current methods have limitations in accounting for sequencing dropouts and retrospective changes in cell state. To address this, we propose LineageCast, a new approach that combines lineage barcode and gene expression data to reconstruct cell division trees more accurately and quickly. We demonstrate LineageCast's effectiveness using data from Caenorhabditis elegans and mouse embryonic cells.

**Session 24CHI81: Recent Advances in Transfer Learning and Domain Generalization**

**Importance Weighted Transfer Learning For High-dimensional Generalized Linear Models**

*Zejing Zheng*1*, Shengbin Zheng*1*，* ♦*Junlong Zhao*1

1Beijing Normal University

Transfer learning o ers a powerful tool for incorporating source data into a target study of interest. Existing methods usually select data at the source level in the sense that data of a given source will be selected or discarded as a whole, requiring that the parameters of an informative source are sufficiently close to that of the target. This requirement might be restrictive. In this paper, we propose a new method named importance weighted transfer learning for high dimensional generalized linear models (IWTL-GLM), selecting data via importance weighting at the individual observation level, thus resulting in more efficient usage of source data. Weights are constructed from conditional density ratios that are estimated in a parametric way, avoiding the curse of dimensionality. The theoretical properties of IWTL-GLM are established and compared with existing methods. Extensive simulations and real data analysis confirm the advantages of our method.

**Structural transfer learning of non-Gaussian DAG with local similarity**

♦*Mingyang Ren*1*, Xin He*2*，Junhui Wang*2

1 The Chinese University of Hong Kong， 2 Shanghai University of Finance and Economics

Directed acyclic graph (DAG) has been widely employed to represent directional relationships among a set of collected nodes. Yet, the available data in one single study is often limited for accurate DAG reconstruction, whereas heterogeneous data may be collected from multiple relevant studies. It remains an open question how to pool the heterogeneous data together for better DAG structure reconstruction in the target study. In this paper, we first introduce a novel set of structural similarity measures for DAG and then present a transfer DAG learning framework by effectively leveraging information from auxiliary DAGs with node-level local similarities. Our theoretical analysis shows substantial improvement in terms of DAG reconstruction in the target study, which is in sharp contrast to most existing transfer learning methods requiring auxiliary DAG with global similarity to the target DAG. The advantage of the proposed transfer DAG learning is also supported by extensive numerical experiments on both synthetic data and multi-site brain functional connectivity network data.

**Knowledge Transfer across Multiple Principal Component Analysis Studies**

*Zeyu Li*1*, Kangxiang Qin*2*,* ♦*Yong He*2*, Wang Zhou*1 *and Xinsheng Zhang*1

1Fudan University，2Shandong University， 3National University of Singapore

Transfer learning has aroused great interest in the statistical community. In this article, we focus on knowledge transfer for unsupervised learning tasks in contrast to the supervised learning tasks in the literature. Given the transferable source populations, we propose a two-step transfer learning algorithm to extract useful information from multiple source principal component analysis (PCA) studies, thereby enhancing estimation accuracy for the target PCA task. In the first step, we integrate the shared subspace information across multiple studies by a proposed method named as Grassmannian barycenter, instead of directly performing PCA on the pooled dataset. The proposed Grassmannian barycenter method enjoys robustness and computational advantages in more general cases. Then the resulting estimator for the shared subspace from the first step is further utilized to estimate the target private subspace in the second step. Our theoretical analysis credits the gain of knowledge transfer between PCA studies to the enlarged eigenvalue gaps of the population covariance matrices, which is different from the existing supervised transfer learning tasks where sparsity of the parameters plays the central role. When the set of informative sources is unknown, we endow our algorithm with the capability of useful dataset selection and solve a rectified optimization problem on the Grassmann manifold, which in turn leads to a computationally friendly rectified Grassmannian K-means procedure. In the end, extensive numerical simulation results and a real data case concerning activity recognition are reported to support our theoretical claims and to illustrate the empirical usefulness of the proposed transfer learning methods.

**Domain Generalization via Low-rank Tensor Completion with Second-Order Convergence**

♦*Taoning Li*1，*Sai Li*1.

1Renmin University of China.

Domain generalization refers to the task of training a model on multiple source domains with different distributions, such that it can generalize well to unseen target domains. However, the most challenging aspect of domain generalization lies in effectively addressing the distribution shift across multiple source domains to enable robust performance on unseen target domains. To address this challenge, we propose a novel tensor-completion based domain generalization algorithm. In the framework of linear regression models, we organize all the regression coefficients within domains into a low-rank tensor. By leveraging the low-rank structure of the tensor, we estimate the regression coefficients on target domains using a tensor completion algorithm. We provide a quadratic convergence guarantee for the algorithm under certain regularity conditions. Finally, numerical simulations are conducted to demonstrate the effectiveness of the proposed approach.

**Session 24CHI52: New Methods on Statistical Learning with Survival Data**

**Nonparametric testing for survival data with time-dependent covariates**

*Ying Cui*1*,* ♦*Limin Peng*1*，HuiChuan Lai*2*.*

1Emory University，2University of Wisconsin-Madison

A time-dependent covariate (e.g., time-varying treatment or exposure) is often encountered in survival studies in biomedical research. The evolving nature of the time-dependent covariate along with the survival outcome poses extra complications in the assessment of the corresponding covariate-outcome association. In this work, we propose a new nonparametric testing framework that is designed to robustly evaluate the effect of a time-dependent covariate on a survival outcome. By adopting the landmark perspective and utilizing a new interval quantile correlation index, our testing procedure does not require parametric or semiparametric modeling of the relationship between the time-dependent covariate and the time-to-event outcome, while flexibly accommodating dynamic covariate effects on the survival outcome. We provide theoretical justifications for our proposals. The new method is applied to probe the effect of time-varying feeding patterns on the pulmonary outcomes of infants with cystic fibrosis.

**Imputation-Based Q-Learning for Optimizing Dynamic Treatment Regimes with Competing Risk Outcomes**

*Lingyun Lyu*1*,* ♦*Yu Cheng*2*， Abdus Wahed* 3

1Bristol Myers Squibb， 2University of Pittsburgh， 3University of Rochester

In this talk, we present an imputation-based Q-learning (IQ-learning) method for optimizing dynamic treatment regimes (DTRs) in multi-stage decision making with competing risk data. A semiparametric cause-specific Cox proportional hazard model is employed to estimate optimal treatment rules for each stage and then weighted hot-deck multiple imputation (MI) and direct-draw MI are used to predict optimal cause-specific potential survival times. We extend the proposed optimal DTR estimation methods to an incomplete-data setting where missing data are handled using inverse probability weighting and MI. We investigate the performance of IQ-learning via extensive simulations and show that it is robust to model misspecification. We demonstrate IQ-Learning by developing an optimal DTR for neuroblastoma treatment based on a randomized trial with observational follow-up where disease progression and death are competing events.

**Estimation of average treatment effect for survival outcomes with continuous treatment in observational studies**

*Triparna Poddar*1*,* ♦*Qi Zheng*1*, Michael Egger*1*， Maiying Kong*1

1University of Louisville

In healthcare research, where extensive observational data such as claims data and electronic records are readily available, researchers often seek to investigate both the treatment effect and the pathway of that effect. While recent literature on causal effects in survival analyses primarily focuses on binary or multiple treatment settings, studies involving continuous treatment settings are rarely explored. In this project, our aim is to explore the estimation of the average treatment effect (ATE) of continuous treatment on time-to-event outcomes by addressing multiple confounding factors and considering censoring observations. We propose to estimate the ATE using the accelerated failure time marginal structural model (AFT-MSM), incorporating the inverse probability of treatment weighting (IPTW) method along with censoring weights. The IPTW method is designed to mitigate the influence of confounding variables on treatment assignment, while censoring weights address potential biases arising from censored observations. Extensive simulation studies have demonstrated the effectiveness of our proposed method. We applied our proposed methodology to investigate the impact of blood lead levels on the time to death among older individuals in the United States, utilizing data from the NHANES III survey dataset.

**Efficient Estimation for the Accelerated Failure Time Model with Auxiliary Aggregate Information**

♦*Huijuan Ma*1*, Yukun Liu, Donglin Zeng and Yong Zhou*

With the rapidly increasing availability of aggregate data in the public domain, there has been a growing interest in synthesizing information from individual-level data and aggregate data. This article studies the maximum full likelihood estimation method to integrate the auxiliary information in the estimation of the accelerated failure time model. To overcome the computational challenges in maximizing full likelihood, we propose a novel one-step estimator, where the maximum conditional likelihood estimator without combining any auxiliary information is chosen as an initial estimator. We establish the consistency and asymptotic normality of the proposed one-step estimator and show that it is more efficient than the initial estimator. The asymptotic variance of the proposed one-step estimator has a closed form and is easily estimated by the plug-in rule. Simulation studies show that the proposed one-step estimator yields an efficiency gain over existing approaches. The proposed methodology is illustrated with an analysis of a chemotherapy study for Stage III colon cancer.

**Session 24CHI93: Recent Developments of Bayesian Methods**

**Enhancing Decision Making with Causal Inference and Unmeasured Confounders in a Bayesian Framework**

*Wei Jin*1, *Yang Ni*2 ， ♦*Yanxun Xu*1*.*

1Johns Hopkins University， 2Taxas A&M University

Navigating the complexities of decision making demands a robust understanding of causal relationships, especially with unmeasured confounders in the background. These hidden variables can obscure accurate assessments. We will talk about a model that integrates hierarchical Bayesian modeling with causal inference methods, unveiling intricate causal connections even in the presence of unobserved variables. The proposed framework not only identifies true causal effects but also equips decision makers to tackle challenges posed by unmeasured confounders. A pivotal step towards informed decisions, this research enhances our ability to comprehend complex systems and harness their insights for effective outcomes.

**Horseshoe Priors for Sparse Dirichlet-Multinomial Models**

♦*Yuexi Wang*1 ，*Nicholas Polson*2*.*

1University of Illinois Urbana-Champaign， 2University of Chicago

Bayesian inference for Dirichlet-Multinomial (DM) models has a long and important history. The concentration parameter $\alpha$ is pivotal in smoothing category probabilities within the multinomial distribution and is crucial for the inference afterward. Due to the lack of a tractable form of its marginal likelihood, $\alpha$ is often chosen in an ad-hoc manner, or estimated using approximation algorithms. A constant $\alpha$ often leads to inadequate smoothing of probabilities, particularly for sparse compositional count datasets. In this paper, we introduce a novel class of prior distributions facilitating conjugate updating of the concentration parameter, allowing for full Bayesian inference for DM models. Our methodology is based on fast residue computation and admits closed-form posterior moments in specific scenarios. Additionally, our prior provides continuous shrinkage with its heavy tail and substantial mass around zero, ensuring adaptability to the sparsity or quasi-sparsity of the data. We demonstrate the usefulness of our approach on both simulated examples and on real-world applications. Finally, we conclude with directions for future research.

**Tree-Regularized Bayesian Latent Class Analysis: Improving Weakly Separated Dietary Pattern Subtyping in Small-Sized Subpopulation**

*Mengbing Li*1, *Briana Stephenson*2 *and* ♦*Zhenke Wu*1

1University of Michigan，2Harvard University

Dietary patterns synthesize multiple related diet components, which can be used by nutrition researchers to examine diet-disease relationships. Latent class models (LCMs) have been used to derive dietary patterns from dietary intake assessment, where each class profile represents the probabilities of exposure to a set of diet components. However, LCM-derived dietary patterns can exhibit strong similarities, or weak separation, resulting in numerical and inferential instabilities that challenge scientific interpretation. This issue is exacerbated in small-sized subpopulations. To address these issues, we provide a simple solution that empowers LCMs to improve dietary pattern estimation. We develop a tree-regularized Bayesian LCM that shares statistical strength between dietary patterns to make better estimates using limited data. This is achieved via a Dirichlet diffusion tree process that specifies a prior distribution for the unknown tree over classes. Dietary patterns that share proximity to one another in the tree are shrunk towards ancestral dietary patterns a priori, with the degree of shrinkage varying across pre-specified food groups. Using dietary intake data from the Hispanic Community Health Study/Study of Latinos, we apply the proposed approach to a sample of 496 US adults of South American ethnic background to identify and compare dietary patterns.

**A Nonstationary Soft Partitioned Gaussian Process Model**

*Zhao Tang Luo*1, ♦*Huiyan Sang*1*， Bani Mallick*1

1Texas A&M University

There has been a long-standing challenge in developing locally stationary Gaussian process models concerning how to obtain flexible partitions and make predictions near boundaries. In this work, we develop a new class of locally stationary stochastic processes, where local partitions are modeled by a soft partition process via predictive random spanning trees that leads to highly flexible spatially contiguous subregion shapes. This valid nonstationary process model knits together local models such that both parameter estimation and prediction can be performed under a unified and coherent framework, and it captures both discontinuities/abrupt changes and local smoothness in a spatial random field. We propose a theoretical framework to study the Bayesian posterior concentration concerning the behavior of this Bayesian nonstationary process model. The performance of the proposed model is illustrated with simulation studies and real data analysis of precipitation rates over the contiguous United States.

**Session 24CHI99: Reinforcement Learning from a Statistical Perspective**

**Statistical- and Communication-Efficient Federated Offline Reinforcement Learning**

♦*Yuejie Chi*1*.*

1Carnegie Mellon University

Offline reinforcement learning (RL), which seeks to learn an optimal policy using offline data, has garnered significant interest due to its potential in critical applications where online data collection is infeasible or expensive. This work explores the benefit of federated learning for offline RL, aiming at collaboratively leveraging offline datasets at multiple agents. Focusing on finite-horizon episodic tabular Markov decision processes (MDPs), we design FedLCB-Q, a variant of the popular model-free Q-learning algorithm tailored for federated offline RL. FedLCB-Q updates local Q-functions at agents with novel learning rate schedules and aggregates them at a central server using importance averaging and a carefully designed pessimistic penalty term. Our sample complexity analysis reveals that, with appropriately chosen parameters and synchronization schedules, FedLCB-Q achieves linear speedup in terms of the number of agents without requiring high-quality datasets at individual agents, as long as the local datasets collectively cover the state-action space visited by the optimal policy, highlighting the power of collaboration in the federated setting. In fact, the sample complexity almost matches that of the single-agent counterpart, as if all the data are stored at a central location, up to polynomial factors of the horizon length. Furthermore, FedLCB-Q is communication-efficient, where the number of communication rounds is only linear with respect to the horizon length up to logarithmic factors.

**Post Reinforcement Learning Inference**

*Vasilis Syrgkanis*1 ， ♦*Ruohan Zhan*2*.*

1Stanford University， 2Hong Kong University of Science and Technology

We consider estimation and inference using data collected from reinforcement learning algorithms. These algorithms, characterized by their adaptive experimentation, interact with individual units over multiple stages, dynamically adjusting their strategies based on previous interactions. Our goal is to evaluate a counterfactual policy post-data collection and estimate structural parameters, like dynamic treatment effects, which can be used for credit assignment and determining the effect of earlier actions on final outcomes. Such parameters of interest can be framed as solutions to moment equations, but not minimizers of a population loss function, leading to Z-estimation approaches for static data. However, in the adaptive data collection environment of reinforcement learning, where algorithms deploy nonstationary behavior policies, standard estimators do not achieve asymptotic normality due to the fluctuating variance. We propose a re-weighted Z-estimation approach with carefully designed adaptive weights to stabilize the time-varying estimation variance. We identify proper weighting schemes to restore the consistency and asymptotic normality of the re-weighted Z-estimators for target parameters, which allows for hypothesis testing and constructing uniform confidence regions. Primary applications include dynamic treatment effect estimation and dynamic off-policy evaluation.

**Sample Complexity of Inventory Control with Fixed Ordering Cost**

♦*Xiaoyu Fan*1*，Zhengyuan Zhou*1.

1New York University.

We show in this work that a class of structured MDPs admits more efficient learning (i.e., lower sample complexity bounds) compared to the best possible/known algorithms in generic RL. We focus on the MDPs describing the inventory control system with fixed ordering costs, a fundamental problem in supply chains. We develop an algorithm applied to the inventory MDPs, which leads to strictly lower sample complexity bounds compared to the optimal or best-known bounds recently obtained for the general MDPs. We improve on those ``best-possible'' bounds by carefully leveraging the structural properties of the inventory dynamics in various settings.

**Settling the Sample Complexity of Model-Based Offline Reinforcement Learning**

♦*Yuting Wei*

University of Pennsylvania

This paper is concerned with offline reinforcement learning (RL), which learns using pre-collected data without further exploration. Effective offline RL would be able to accommodate distribution shift and limited data coverage. However, prior algorithms or analyses either suffer from suboptimal sample complexities or incur high burn-in cost to reach sample optimality, thus posing an impediment to efficient offline RL in sample-starved applications. We demonstrate that the model-based (or "plug-in") approach achieves minimax-optimal sample complexity without burn-in cost for tabular Markov decision processes (MDPs). We prove that model-based offline RL yields epsilon-accuracy with a sample complexity that is minimax optimal for the entire epsilon-range. The proposed algorithms are ``pessimistic'' variants of value iteration with Bernstein-style penalties, and do not require sophisticated variance reduction. Our analysis framework is established upon delicate leave-one-out decoupling arguments in conjunction with careful self-bounding techniques tailored to MDPs.

**Session 24CHI119: Statistical Methods and Applications for Healthy Ageing**

**Examining the impact of demographic factors on disability in Chinese elderly using a semiparametric copula model**

♦*Tao Sun*1, *Yunlong Li*1, *Zhengyan Xiao*1, *Ying Ding*2 *， Xiaojun Wang*1.

1Renmin University of China 2University of Pittsburgh

According to the World Health Organization (WHO), over 46% of individuals aged 60 and above experience disabilities. Assessing the influence of demographic factors on time-to-disability in the elderly is crucial for early prevention and management. A significant challenge in studying time-to-disability in the elderly is the high mortality rate, leading to semi-competing risks data where death censors disability but not vice versa. Copula models offer an approach to address this, explicitly considering the dependence between disability and death. However, a gap exists in fitting such a model for disability data due to two challenges: interval censoring due to intermittent assessments and left truncation when time is on the age scale. We present a novel two-parameter copula-based semiparametric transformation model that handles interval censoring and left truncation in semi-competing risks data. This copula quantifies both upper and lower tail dependence between the marginal distributions of disability and death. Semiparametric transformation models incorporate various assumptions in both margins. Numerical simulations demonstrate the proposed method's ability to correct estimation biases and provide accurate coverage probabilities. Applied to a dataset of 12,969 individuals over 60 from the Chinese Longitudinal Healthy Longevity Survey, our method assesses the impact of demographic factors (age, sex, education, marriage) on disability and death. This research is the first to use a robust statistical model for disability, addressing interval censoring and left truncation in semi-competing risks data, providing new insights into elderly disability prevention and management.

**Polygenic risk score for disability development in elderly**

♦*Huiping Zheng*1, *Tao Sun*1*， Xiaojun Wang*1.

1Renmin University of China

Elderly disability results from multiple factors and has significant impacts on the long-term care demand as well as social and economic development. However, current research on the genetic mechanisms underlying disability is limited, and their influence on the development of disability remains unclear. This study aims to identify genetic variants associated with disability using genomic and questionnaire data from the Canadian Longitudinal Study on Aging (CLSA). We firstly conduct a genome-wide association analysis involving 8,421 individuals aged 65 and older. Specifically, we use the Cox proportional hazards model to examine the effect of 8 million single nucleotide polymorphisms (SNPs) on the development of disability. Our study identifies numerous significant disability-associated SNPs. Subsequently, we derived a polygenic risk score (PRS) from these risk variants and evaluated its impact on the development of disability. Our results provide new insights into the genetics of elderly disability.

**A Copula-based Semiparametric Model for Predicting Life Expectancy and Healthy Life Expectancy**

♦*Yunlong Li*1, *Xiaojun Wang*1 *，Tao Sun*1

1Renmin University of China

Population ageing is a global phenomenon with significant implications for both developed and developing countries. Understanding the dynamics of healthy life expectancy and life expectancy is crucial for policy makers, healthcare providers and insurers. However, accurate predictions of healthy life expectancy and life expectancy are limited by censored and truncated data from longitudinal surveys.

In this study, we present a new method for calculating healthy life expectancy and life expectancy based on a joint disability-death survival model. We use the two-parameter copula model to capture the relationship between death and disability that has been overlooked in previous models. We constructed semi-parametric transformation models as marginal models and incorporated right censoring, interval censoring, left censoring and left truncation into the likelihood function to ensure that our model adequately handles the complex nature of longitudinal data. In addition, our model incorporates various covariates, including gender, location, and education level, which allows the calculation of life expectancy and healthy life expectancy for populations with different characteristics.

Through simulations, we show that our model has better estimation performance in the presence of incomplete data compared to existing multi-state Markov models. Using data from the Chinese Longitudinal Healthy Longevity Survey (CLHLS), we empirically analyze the prediction of life expectancy and healthy life expectancy in China up to 2030.

**The difference of health-adjusted life expectancy between China and both the United States and Japan**

♦*Yueming Jin*1*， Xiaojun Wang*1

1Renmin University of China

As the aging of the population accelerates and life expectancy increases, the elderly live longer with illness or disability. In this context, healthy life expectancy, which measures both length and quality of life of a population, has become an important monitoring indicator for building a healthy country. So, how big is the gap in healthy life expectancy between China and other developed countries? What are the main reasons for these gaps? How to better improve the quality of life of our country's population? To answer these questions, this study selects the United States and Japan, two developed countries that have earlier included healthy life expectancy monitoring indicators in their national health plans, for comparative research. In terms of research methods, the study expands the Andreev decomposition method based on the stepwise replacement algorithm, based on the Global Burden of Disease database (GBD2019) and the Sullivan method, compares the healthy life expectancy gap between both China and the United States and China and Japan, and decomposes the health gap between countries into the death effects and disability effects of different groups of age and different causes of diseases or injuries, which can be further used to derive the subdivided influencing factors of health disparities. The results show that the gap in healthy life expectancy between China and the United States and Japan is mainly reflected in the relatively high mortality rate of infants and people over 70 years old in our country. Compared with the United States and Japan, China's non-communicable chronic diseases, especially cardiovascular diseases, chronic obstructive pulmonary disease, and certain malignant tumors have high mortality rates, while the mortality rates from infectious diseases are relatively low. For example, the death effect of stroke is the largest contributor to the differences in healthy life expectancy between both China and the United States and China and Japan, with gaps of 1.22 and 1.55 years respectively. Also, deaths caused by chronic obstructive pulmonary disease make China's healthy life expectancy lower than that of the United States and Japan by 0.29 and 0.59 years respectively. Although the overall contribution of the death effect of injuries to health disparities is small, it is found specifically that China performs better in terms of self-harm and interpersonal violence, while accidental injuries such as drowning and falls have higher mortality rates. It is recommended that in the building of a healthy China, we should focus on the prevention and treatment of major chronic diseases, further reduce infant and child mortality, and promote the improvement of both population life expectancy and healthy life expectancy.

**Session 24CHI7: Advanced Statistical Methods in the Analysis of High-Dimensional and Complex Data**

**Ensemble parameter-transfer by optimal model averaging**

♦*Xiaonan Hu*1*，Baqun Zhang*2.

1Capital Normal University， 2Chinese Academy of Sciences

Transfer learning, as a prevalent technique in computer science, provides a powerful framework for the analysis of data from multiple sources. This talk introduces a novel parameter-transfer approach for high-dimensional generalized linear models. To tackle the potential uncertainty of auxiliary information in applications, we propose an ensemble method that leverages a frequentist model averaging strategy. We develop a KL-type weight choice criterion based on a J-fold cross-validation procedure and establish the statistical properties in both misspecified and correctly specified scenarios of the target model. Extensive numerical results demonstrate the superiority of our method over competing alternatives.

**Network and covariate adjusted response-adaptive design**

♦*Hao Mei*1*, Jiaxin Xie*1*, Jiaxin Xie*2， *Yichen Qin*1*.*

1Renmin University of China, 2University of Cincinnati

Randomization is a distinguishing feature of clinical trials for unbiased assessment of treatment efficacy. With a growing demand for more flexible and efficient randomization schemes and motivated by the idea of adaptive design, in this article we propose the network and covariate adjusted response-adaptive (NCARA) design that can concurrently manage three challenges: (1) maximizing benefits of a trial by assigning more patients to the superior treatment group randomly; (2) balancing social network ties across treatment arms to eliminate potential network interference; and (3) ensuring balance of important covariates, such as age, gender, and other potential confounders. We conduct simulation with different network structures and a variety of parameter settings. It is observed that the NCARA design outperforms four alternative randomization designs in solving the above-mentioned problems and has comparable power and type I error for detecting true difference between treatment groups. In addition, we conduct real data analysis to implement the new design in two clinical trials. Compared to equal randomization (the original design utilized in the trials), the NCARA design slightly increases power, largely increases the percentage of patients assigned to the better-performing group, and significantly improves network and covariate balances. It is also noted that the advantages of the NCARA design are augmented when the sample size is small and the level of network interference is high. In summary, the proposed NCARA design assists researchers in conducting clinical trials with high-quality and high-efficiency.

**Hierarchical Multi-Label Classification with Gene-Environment Interactions in Disease Modeling**

*Jingmao Li1, Qingzhao Zhang*1*, Shuangge Ma*2*, Kuangnan Fang*1， *♦Yaqing Xu3*

1Xiamen University， 2Yale University， 3Shanghai Jiao Tong University

In biomedical studies, gene-environment (G--E) interactions have been demonstrated to have important implications for disease prognosis beyond the main G and main E effects. Many approaches have been developed for G-E interaction analysis, yielding important findings. However, hierarchical multi-label classification, which provides insightful information on disease outcomes, remains unexplored in G-E analysis literature. Moreover, unlabeled data is commonly observed in practical settings but omitted by many existing methods of hierarchical multi-label classification. In this study, we consider a semi-supervised scenario and develop a novel approach for the two-layer hierarchical response with G-E interactions. A two-step penalized estimation is then proposed using an efficient expectation-maximization (EM) algorithm. Simulation shows that it has superior accuracy in classification and feature selection performance. The analysis of The Cancer Genome Atlas (TCGA) data on lung cancer demonstrates the practical utility of the proposed approach. Overall, this study can fill the important knowledge gap in G-E interaction analysis by providing a widely applicable framework for hierarchical multi-label classification for complex disease outcomes.

**Deep learning-based single-cell data integration using iterative cell matching and structure preservation constrains**

*Shuntuo Xu1, Zhou Yu1 , ♦Jingsi Ming1*

1East China Normal University

Effective integration of single-cell data can facilitate the discovery of cell-type specific gene expression patterns and cellular interactions, ultimately leading to a better understanding of various biological processes and diseases. However, datasets from different platforms, species, and modalities exhibit various levels of heterogeneities, posing significant challenges in data alignment using a unified approach. Here we propose DeepMap, a flexible and efficient method for single-cell data integration, by taking advantage of the deep learning framework. Our method utilizes iterative cell matching based on mutual nearest neighbors, leverages an autoencoder framework to learn harmonized representations of cells from various datasets, and incorporates a covariance penalty term into the framework for structure preservation. In addition to harmonization of data from different datasets, we specifically take account of the preservation of important biological variations within dataset, which is crucial to reliable downstream analysis. Comprehensive real data analysis demonstrates the flexibility of DeepMap for diverse datasets from different platforms, species and modalities, and highlights its marked ability in preserving structures over existing integration methods with enhanced computational efficiency and optimized memory usage. The robust DeepMap-integrated data offers promising prospects for advancing our understanding of cell biology, hence making it a highly attractive option for integrative single-cell data analysis.

**Session 24CHI33: Go No Go Decision Making in Clinical Trails**

**Go/No go design based on dual criterion**

♦*Chao Cheng*1*, Haolin Sun*1*, Jonathan Haft*1 *，Robert Hawkins*1*.*

1Simcere Zaiming Pharmaceutical Co, Ltd.

Proof of concept (POC) is a milestone in drug development. Typically, after formal evaluation one of the following decisions will be made: continue further development (Go), stop development (Nogo), or seek further information. In POC trials, dual-criterion design can be applied to investigate the drug efficacy. The decision criteria are two-fold: 1. whether the drug is superior than reference/placebo, 2. whether the improvement is clinically meaningful. In this talk I will introduce the principles of dual-criterion design and demonstrate its application with different types of endpoints via case studies.

**From Phase II to III: An Application of Success Probability Estimation in Oncology**

♦*Ligiang Geng*1.

1Boehringer Ingelheim.

Despite the escalating number of drugs under development, the success rate from Phase III trials to final drug approval is dwindling. Consequently, leveraging quantitative methods to guide pivotal decisions, such as advancing a drug after the Phase II study, is becoming increasingly crucial. One such method is to estimate the Probability of Success (PoS), which factors in the variability of random observation to yield a more realistic success probability than power calculation. The presentation will focus on an application in oncology where overall survival (OS) is considered as the primary endpoint for PhIII study. It will illustrate how the PoS for OS, based on Phase II results, can aid in the Go/No-Go decision for Phase III entry. Given the dependence of Progression-Free Survival (PFS) and OS, we will also investigate a multi-state model that includes the statuses of progression and death to jointly model PFS and OS.

**Bayesian method used in go no-go decision**

♦*Hairui Hua*1

In early clinical development of a new treatment, it is critical to establish the proof of concept (PoC) in phase II studies. The PoC studies typically pre-define a “Go/NoGo” decision rule to help the clinical researchers to assess the performance of the treatment of interests. If the observed efficacy data meets the “Go” decision bar, it could reach the conclusion that the concept has been proven and the further development such as Phase IIb or even pivotal Phase III studies with more patients and longer follow up period will be initiated.

Generally, in PoC phase, it is not ethical to expose too many patients under the treatment of which the effects are still unknown. Further the clinical researchers are not willing to spending too much money at this stage, either. Therefore, the sample size of PoC studies are generally low and often it may be hard to make a clear Go-NoGo decision with the limited data alone. To address this issue, Bayesian borrowing methods have been introduced. It allows us to include historical data into both active and control arms which provide the researcher more power to make the correct Go-NoGo decision. This presentation will talk about the statistical ideas behind Bayesian borrowing method and meanwhile share one or two related industrial cases.

**RESTART trial design: two-stage seamless transition design with operational considerations**

*Min Yi*1, ♦*Bin Zhuo*2 ，*Freda Cooner*3.

Oncology/hematology is a competitive therapeutic area where the landscape is constantly evolving. With regulatory support, many drug developers have spent a lot of resources on the operationalization of innovative clinical trial designs, for example, adaptive Bayesian designs in confirmatory clinical trial settings. While overall survival is considered the gold standard in these

designs, it is often not a viable choice in identifying treatment efficacy at a reasonable pace, especially for early-stage therapies. In recent years, several binary response surrogate endpoints have been used for accelerated or conditional approval of novel cancer therapies. Utilizing surrogate endpoints in the study design to predict objective clinical outcomes, such as overall survival, is particularly fundamental in cancer treatment clinical development. This manuscript will investigate logistic and statistical considerations of our proposed RESTART design, a new two-stage, seamless, single- to double-arm Bayesian design. This design could be used for single-arm dose expansion to a randomized confirmatory study. The operating characteristics of the RESTART design are evaluated based on simulations. Future directions and further modifications of this design will also be elaborated.

**Session 24CHI38: Inference for Paired Comparision Data and Clustering**

**Heteroskedastic Tensor Clustering**

*Yuchen Zhou*1*，* ♦*Yuxin Chen*1*.*

1University of Pennsylvania

Tensor clustering, which seeks to extract underlying cluster structures from noisy tensor observations, has gained increasing attention. One extensively studied model for tensor clustering is the tensor block model, which postulates the existence of clustering structures along each mode and has found broad applications in areas like multi-tissue gene expression analysis and multilayer network analysis. However, currently available computationally feasible methods for tensor clustering either are limited to handling i.i.d. sub-Gaussian noise or suffer from suboptimal statistical performance, which restrains their utility in applications that have to deal with heteroskedastic data and/or low signal-to-noise-ratio (SNR).

To overcome these challenges, we propose a two-stage method, named 𝖧𝗂𝗀𝗁-𝗈𝗋𝖽𝖾𝗋 𝖧𝖾𝗍𝖾𝗋𝗈𝖢𝗅𝗎𝗌𝗍𝖾𝗋𝗂𝗇𝗀 (𝖧𝖧𝖢), which starts by performing tensor subspace estimation via a novel spectral algorithm called 𝖳𝗁𝗋𝖾𝗌𝗁𝗈𝗅𝖽𝖾𝖽 𝖣𝖾𝖿𝗅𝖺𝗍𝖾𝖽-𝖧𝖾𝗍𝖾𝗋𝗈𝖯𝖢𝖠, followed by approximate k-means to obtain cluster nodes. Encouragingly, our algorithm provably achieves exact clustering as long as the SNR exceeds the computational limit (ignoring logarithmic factors); here, the SNR refers to the ratio of the pairwise disparity between nodes to the noise level, and the computational limit indicates the lowest SNR that enables exact clustering with polynomial runtime. Comprehensive simulation and real-data experiments suggest that our algorithm outperforms existing algorithms across various settings, delivering more reliable clustering performance.

**Inference in a generalized Bradley-Terry model for paired comparisons with covariates and a growing number of subjects**

♦*Ting Yan*1

1Central China Normal University

The Bradley--Terry model has been widely used to rank subjects participating in paired comparisons Asymptotic theories in this model have been revealed when the number of subjects grows. However, little is still known when covariate information (e.g., home-field advantage) appears. In this paper, we propose a generalized Bradley-Terry model by incorporating the covaraites. When the number of subjects $n$ goes to infinity and the number of comparisons between any two subjects is fixed, we show that the respective upper bounds of the errors $\| \widehat{\beta} - \beta\|\_\infty$ and $\| \widehat{\gamma} - \gamma\|\_\infty$ are $O(n^{-1/2})$ and $O(n^{-1})$, both up to a logarithm factor, where $\beta$ is the merit parameter vector, $\gamma$ is the covariate parameter and $(\widehat{\beta}, \widehat{\gamma})$ is the maximum likelihood estimator (MLE) of $(\beta, \gamma)$. Further, we derive the asymptotic normal distribution of the MLE by characterizing its asymptotic representation. This is proved by applying Taylor's expansions to a series of functions constructed from likelihood equations and showing remainder terms in the expansions are asymptotically neglect. To the best of our knowledge, this is the first time to explore asymptotic theory in paired comparison models with covariates in the high dimensional setting. Our paper sheds light on how to explore asymptotic theory in the appearance of the covaraites in a principled manner; these principled methods should be applicable to a class of paired comparison models beyond the generalized Bradley-Terry model.

**Structure Recognition for Time-varying Comparisons**

♦*Nan Lu*1, *Jian Shi*1*， Xin Yutian*2.

1Academy of Mathematics and Systems Science， 2University of Minnesota.

Dynamic ranking of numerous items often yields complex outcomes, making it challenging to convey intuitive impressions. In populations, latent structures with unobserved clustering frequently exist, indicating homogeneous items that exhibit similar behavior. Furthermore, the grouping of individuals may change over time, necessitating the detection of structural change points for long-term analysis. We propose a novel method for recognizing structures in dynamic ranking. We first introduce an approach for simultaneously grouping and ranking within a specified time interval. Subsequently, we present a combined penalty for structure change detection and propose a practical algorithm based on dynamic programming. Theoretically, we leverage the properties of a random 'design matrix' induced by a reversible Markov chain to prove the grouping consistency and deduce the asymptotic distribution of estimators. We also develop a theory on change point estimation accuracy, ensuring robust and reliable results. To demonstrate the effectiveness of our method, we apply it to both synthetic data and real datasets.

**Statistical inference for pairwise comparison models**

♦*Ruijian Han*1, *Wenlu Tang*2 ， *Yiming Xu*3.

1The Hong Kong Polytechnic University， 2University of Alberta and 3University of Waterloo.

Pairwise comparison models have been widely used for utility evaluation and ranking across various fields. The increasing scale of problems today underscores the need to understand statistical inference in these models when the number of subjects diverges, a topic currently lacking in the literature except in a few special instances. To partially address this gap, this paper establishes a near-optimal asymptotic normality result for the maximum likelihood estimator in a broad class of pairwise comparison models, as well as a non-asymptotic convergence rate for each individual subject under comparison. The key idea lies in identifying the Fisher information matrix as a weighted graph Laplacian, which can be studied via a meticulous spectral analysis. Our findings provide a unified theory for performing statistical inference in a wide range of pairwise comparison models beyond the Bradley--Terry model, benefiting practitioners with theoretical guarantees for their use. Simulations utilizing synthetic data are conducted to validate the asymptotic normality result, followed by a hypothesis test using a tennis competition dataset.

**Session 24CHI90: Recent Developments in Longitudinal Data Analysis**

**Multivariate Dependence Modeling of Some Data Sets in Continuous Space and Discrete Time**

♦*Juan Du*1*， Rigele Te* 1

1Kansas State University

Space- time data are often multivariate and collected at monitored discrete time lags, which are usually viewed as a component of time series. Valid and practical covariance models are needed to capture the complex dependence structure of these types of data sets in various disciplines, such as environmental science, climatology, and agriculture. We propose and characterize several classes of multivariate spatio-temporal covariance matrix functions whose temporal margins are some celebrated autoregressive and moving average models. This model specification enables us to take advantage of well-established time series and spatial statistics tools, which make the model identification relatively straightforward in practice. The applications of proposed multivariate covariance matrix models are illustrated using simulation study and Kansas weather data in terms of co-kriging, compared with some traditional space-time models for prediction.

**A functional correlation based model-free feature screening method for ultrahigh dimensional longitudinal data**

*Yexin Zhang*1*，* ♦*Shirong Deng*1

1Wuhan University

For longitudinal data with ultrahigh dimensional time-dependent covariates, a model-free feature screening method is proposed, which ranks the importance of the time-dependent covariates according to their new defined marginal functional correlations with the response process. The theoretical properties of our proposed functional correlation between two stochastic processes are given. Furthermore, the sure independence screening property, ranking consistency and false selection rate control of the corresponding feature screening procedure based on new functional correlation are well established. Extensive simulation studies and a real data analysis based on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset demonstrate the superior and promising screening performance of our method.

**An Optimal Two-Step Estimation Approach for Two-Phase Studies**

♦*Kin Yau Wong*1， *Qingning Zhou*2

1The Hong Kong Polytechnic University， 2The University of North Carolina at Charlotte

Two-phase sampling is commonly adopted for reducing cost and improving estimation efficiency. In this project, we consider the two-phase study design where the outcome and some cheap covariates are observed for a large cohort at Phase I, and expensive covariates are obtained for a selected subset of the cohort at Phase II. As a result, the analysis of the association between the outcome and covariates faces a missing data problem. The complete case analysis that uses only the Phase II sample is generally inefficient. We develop a two-step estimation approach, which first obtains an estimator using the complete data, and then updates it using an asymptotically mean-zero estimator obtained from a working model between the outcome and cheap covariates using the full data. The two-step estimator is asymptotically at least as efficient as the complete-data estimator and is robust to misspecification of the working model. We propose a kernel-based method to construct a two-step estimator that achieves optimal efficiency, and also develop a simple joint update approach based on multiple working models to approximate the optimal estimator. The proposed method is based on the influence function and is generally applicable as long as the complete-data estimator is asymptotically linear. We demonstrate the advantages of the proposed method over the existing approaches via simulation studies and provide applications to real biomedical studies.

**Scalable regression analysis for non-normal longitudinal data in the presence of informative observation times with application to EHR data**

♦*Dayu Sun*

Indiana University

The merging EHR data has the potential to provide rich information for individualized treatment. Nevertheless, EHR data have been posing challenges to traditional longitudinal analysis. The responses in EHR data tend to be skewed and heavy-tailed, leading to biased inference and prediction for conventional methods with normality assumption. Furthermore, the record times for subjects in EHR data are usually correlated to the subject conditions, leading to informative observation times. To our knowledge, limited work has addressed the non-normality and informative observation times simultaneously in the small-scale longitudinal data. Moreover, these existing methods are not computationally scalable to the large size of EHR data. We proposed an efficient and scalable EM algorithm for joint regression modeling for the longitudinal and observation process with multivariate skewed t-distribution. Extensive simulation studies have justified the proposed method and demonstrated its scalability. We illustrated our proposal by application to large dental EHR data with more than 20000 subjects.

**Session 24CHI91: Recent Developments in Machine Learning Theory and Application**

**Estimating heterogeneous survivor causal effects using Bayesian machine learning**

♦*Guangyu Tong*1*, Xinyuan Chen*2*, Michael Harhay*3*， Fan Li*1

1Yale University，2Mississippi State University， 3University of Pennsylvania

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 specify separate mean models for the potential outcomes and latent stratum membership flexibly. 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.

**Generalization and Risk Bounds for Recurrent Neural Networks**

♦*Xuewei Cheng*1*, Ke Huang*2*， Shujie Ma*2.

1Hunan Normal University， 2University of California

Recurrent Neural Networks (RNNs) have achieved great success in the prediction of sequential data. However, their theoretical studies are still lagging behind because of their complex interconnected structures. In this paper, we establish a new generalization error bound for vanilla RNNs, and provide a unified framework to calculate the Rademacher complexity that can be applied to a variety of loss functions. When the ramp loss is used, we show that our bound is tighter than the existing bounds based on the same assumption on the Frobenius and spectral norms of the weight matrices and a few mild conditions. Moreover, we derive a sharp estimation error bound for RNN-based estimators obtained through empirical risk minimization (ERM) in multi-class classification problems when the loss function satisfies a Bernstein condition.

**Unraveling Projection Heads in Contrastive Learning: Insights from Expansion and Shrinkage**

*Yu Gui*1*,* ♦*Cong Ma*1*， Yiqiao Zhong*2.

1University of Chicago， 2University of Wisconsin-Madison

We investigate the role of projection heads, also known as projectors, within the encoder-projector framework (e.g., SimCLR) used in contrastive learning. We aim to demystify the observed phenomenon where representations learned before projectors outperform those learned after---measured using the downstream linear classification accuracy, even when the projectors themselves are linear. In this paper, we make two significant contributions towards this aim. Firstly, through empirical and theoretical analysis, we identify two crucial effects---expansion and shrinkage---induced by the contrastive loss on the projectors. In essence, contrastive loss either expands or shrinks the signal direction in the representations learned by an encoder, depending on factors such as the augmentation strength, the temperature used in contrastive loss, etc. Secondly, drawing inspiration from the expansion and shrinkage phenomenon, we propose a family of linear transformations to accurately model the projector’s behavior. This enables us to precisely characterize the downstream linear classification accuracy in the high-dimensional asymptotic limit. Our findings reveal that linear projectors operating in the shrinkage (or expansion) regime hinder (or improve) the downstream classification accuracy. This provides the first theoretical explanation as to why (linear) projectors impact the downstream performance of learned representations. Our theoretical findings are further corroborated by extensive experiments on both synthetic data and real image data.

**Averaged transfer learning for high-dimensional generalized linear models with auxiliary model uncertainty**

♦*Xinyu Zhang*1.

1Academy of Mathematics and Systems Science Chinese Academy of Sciences.

High dimensionality and multi-source heterogeneity are important characteristics of big data. Transfer learning, as a prevailing technique in computer sciences, provides a powerful framework for analyzing multi-source data. In this article, we propose a parameter-transfer approach for high-dimensional generalized linear models. To address the potential uncertainty of auxiliary information in practice, we introduce an averaging transfer learning procedure (ATL) that borrows the idea of frequentist model averaging based on the Kullback–Leibler (KL) loss. We establish several statistical properties, including asymptotic optimality, weight convergence, and estimation consistency, in cases where the target model is either misspecified or correctly specified. Extensive numerical results demonstrate the advantages of our method compared to alternatives under various settings.

**Session 24CHI61: Recent Advances in Bayesian Modeling of Spatially Resolved Transcriptomics Data**

**AI-powered Bayesian methods for spatial transcriptomics data analysis**

♦*Qiwei Li*1

1The University of Texas at Dallas

Recent technology breakthroughs in spatially resolved transcriptomics have enabled the comprehensive molecular characterization of single cells while preserving their spatial and morphological contexts. This new bioinformatics scenario advances our understanding of molecular and cellular spatial organizations in tissues, fueling the next generation of scientific discovery. Bayesian statistics relies more on human analyses with computer aids, while AI relies more on computer algorithms with aids from humans. This talk will outline methodologies for synergizing AI capabilities with Bayesian frameworks, aiming to resolve key issues in this emerging field. Particularly, I will demonstrate how to integrate information from AI tools into Bayesian models and Bayesian nonparametric approaches to enhance spatial domain identification and enable gene expression reconstruction at the single-cell level.

**Bayesian Integrative Region Segmentation in Spatially Resolved Transcriptomic Studies**

*Yinqiao Yan*1*，* ♦*Xiangyu Luo*1.

1Renmin University of China

The spatially resolved transcriptomic study is a recently developed biological experiment that can measure gene expressions and retain spatial information simultaneously, opening a new avenue to characterize fine-grained tissue structures. We propose a nonparametric Bayesian method named BINRES to carry out the region segmentation for a tissue section by integrating all the three types of data generated during the study—gene expressions, spatial coordinates, and the histology image. BINRES is able to capture more subtle regions than existing statistical partitioning models that only partially make use of the three data modes and is more interpretable than neural-network-based region segmentation approaches. Specifically, due to a nonparametric spatial prior, BINRES does not require a prespecified region number and can learn it automatically. BINRES also combines the image and the gene expressions in the Bayesian consensus clustering framework and thus flexibly adjusts their label alignment contribution weights in a data-adaptive manner. A computationally scalable extension is developed for large-scale studies. Both simulation studies and the real application to three mouse spatial transcriptomic datasets demonstrate that BINRES outperforms the competing methods and easily achieves the uncertainty quantification of the integrative partition.

**Generalized Bayesian nonparametric clustering framework for spatial transcriptomics data**

♦*Bencong Zhu*1*, Xiaodan Fan*1*， Qiwei Li*2

1The Chinese University of Hong Kong， 2The University of Texas at Dallas

The emergence of next-generation sequencing-based spatially resolved transcriptomics (SRT) techniques has revolutionized genomic research by enabling high-throughput gene expression profiling while preserving spatial information. However, when it comes to clustering analysis of these novel high-dimensional spatial data, most existing methods adopt a two-step approach. They first reduce the dimensionality of the expression matrix and then perform model-based clustering analysis on the resulting low-dimensional embeddings. Additionally, in real-world applications of SRT data, the number of underlying groups or clusters is often unknown and needs to be inferred from the data. Consequently, there is a lack of a comprehensive clustering framework that directly utilizes the high-dimensional SRT data while automatically inferring the number of clusters. To address these limitations, we propose a novel class of Bayesian nonparametric mixture of factor analysis (BNPMFA) models. These models leverage a Markov random field (MRF)-constrained Gibbs-type prior on the partition process to infer sample clusters in SRT data. The new prior incorporates the spatial constraints inherent in SRT data while simultaneously inferring cluster membership and determining the number of clusters. We have mathematically established the identifiability of cluster membership within this framework. The performance of our proposed approach is demonstrated through realistic simulations and applied to two SRT datasets. The results highlight that our method outperforms state-of-the-art alternatives in terms of clustering accuracy.

**Session 24CHI59: Recent Advance in Analysis of Network and Matrix Time Series Data**

**Matrix Quantile Factor Model**

*Xin-Bing Kong1, Yong-Xin Liu1,* ♦*Long Yu2, Peng Zhao3*

1Nanjing Audit University，2Shanghai University of Finance and Economics， 3Jiangsu Normal University

This paper introduces a matrix quantile factor model for matrix-valued data with a low- rank structure. We estimate the row and column factor spaces via minimizing the empirical check loss function over all panels. We show the estimates converge at rate 1/ min{√p1p2, √p2T , √p1T} in average Frobenius norm, where p1, p2 and T are the row dimensionality, column dimensionality and length of the matrix sequence. This rate is faster than that of the quantile estimates via “flattening” the matrix model into a large vector model. Smoothed estimates are given and their central limit theorems are derived under some mild condition. We provide three consistent criteria to determine the pair of row and column factor numbers. Extensive simulation studies and an empirical study justify our theory.

**Supervised Centrality via Sparse Network Influence Regression**

♦*Yingying Ma*, *Wei Lan*, *Chenlei Leng*, *Ting Li*, *Hansheng Wang*

The social characteristics of players in a social network are closely associated with their network positions and relational importance. Identifying those influential players in a network is of great importance as it helps to understand how ties are formed, how information is propagated, and in turn can guide the dissemination of new information. Motivated by a Sina Weibo social network on 2021 Henan Floods where response variables on each node are available, we propose a new notion of supervised centrality emphasizing the fact that the centrality of a player is task-specific. To estimate the supervised centrality and identify important players, we develop a novel sparse spatial autoregression by introducing individual heterogeneity to each user. To overcome the computational difficulties in fitting the model for large social networks, we further develop a forward-addition algorithm and show that it can consistently identify a superset of the influential nodes. We apply our method to analyze three responses in Henan Floods data: the number of comments, reposts and likes, and obtain meaningful results. Simulation study further corroborates the developed theory.

**Auxiliary Learning and its Statistical Understanding**

♦*Hanchao Yan*1, *Feifei Wang*2, *Chuanxin* *Xia*3, *Hansheng Wang*4

1School of Statistics, Renmin University of China, Beijing, China, 2Center for Applied Statistics, Renmin University of China, Beijing, China; School of Statistics, Renmin University of China, Beijing, China, 3Chinese Academy of International Trade and Economic Cooperation, Ministry of Commerce People's Republic of China, Beijing, China, 4Guanghua School of Management, Peking University, Beijing, China

Modern statistical analysis often encounters high-dimensional problems but with a limited sample size. It poses great challenges to traditional statistical estimation methods. To address this issue, various statistical methods have been developed, most of which focus on penalized estimation and feature screening. In the field of machine learning, the method of auxiliary learning is demonstrated to be a useful alternative in many applications. The key idea of auxiliary learning is introducing a set of auxiliary learning tasks to enhance the performance of the primary learning task. In this work, we adopt auxiliary learning to solve the estimation problem in high-dimensional settings. We start with the linear regression setup. To improve the statistical efficiency of the parameter estimator for the primary task, we consider several auxiliary tasks, which share the same covariates with the primary task. Then a weighted estimator for the primary task is developed, which is a linear combination of the ordinary least squares estimators of both the primary task and auxiliary tasks. The optimal weight is analytically derived and the statistical properties of the corresponding weighted estimator are studied. We then extend the weighted estimator to generalized linear regression models. Extensive numerical experiments are conducted to verify our theoretical results. Last, a deep learning-related real-data example of smart vending machines is presented for illustration purposes.

**Network-Based Modeling of Emotional Expressions for Multiple Cancers Via a Linguistic Analysis of an Online Health Community**

♦*Xinyan Fan*

The diagnosis and treatment of cancer can evoke a variety of adverse emo- tions. Online health communities (OHCs) provide a safe platform for cancer patients and those closely related to express emotions without fear of judgement or stigma. In the literature, linguistic analysis of OHCs is usually limited to a single disease and based on methods with various technical limitations. In this article, we analyze posts from September 2003 to September 2022 on eight cancers that are publicly available at the American Cancer Society’s Cancer Survivors Network (CSN). We propose a novel network analysis technique based on low rank matrices. The proposed approach decomposes the emotional expression semantic networks into an across-cancer time-independent component (which describes the “baseline” that is shared by multiple cancers), a cancer-specific time-independent component (which describes cancer-specific properties), and an across-cancer time-dependent component (which accommodates temporal effects on multiple cancer com- munities). For the second and third components, respectively, we consider a novel clustering structure and a change point structure. A penalization approach is developed, and its theoretical and computational properties are carefully established. The analysis of the CSN data leads to sensible networks and deeper insights into emotions for cancer overall and specific cancer types.

**Session 24CHI77: Recent Advances in Statistical Methods for Complex Time-To-Event Data**

**Enhancing Long-term Survival Prediction with Multiple Short-term Events: Landmarking with A Flexible Varying Coefficient Model**

♦*Wen Li1, Qian Wang1, Jing Ning2, Jing Zhang1, Zhouxuan Li1, Amirali Tahanan1 ，Mohammad H. Rahbar1.*

1The University of Texas Health Science Center， 2University of Texas MD Anderson Cancer Center

Patients with cardiovascular diseases who experience disease-related short-term events, such as hospitalizations, often exhibit diverse long-term survival outcomes compared to others. In this study, we aim to improve the prediction of long-term survival probability by incorporating multiple short-term events using a flexible varying coefficient landmark model. Our objective is to predict the long-term survival among patients who survived up to a pre-specified landmark time since the initial admission. Inverse probability weighting estimation equations are formed based on the information of the short-term outcomes before the landmark time. The kernel smoothing method with the use of cross-validation for bandwidth selection is employed to estimate the time-varying coefficients. The predictive performance of the proposed model is evaluated and compared using predictive measures: area under the receiver operating characteristic curve and Brier score. Simulation studies confirm that parameters under the landmark models can be estimated accurately and the predictive performance of the proposed method consistently outperforms existing methods that either do not incorporate or only partially incorporate information from two short-term events. We demonstrate the practical application of our model using a community-based cohort from the Atherosclerosis Risk in Communities (ARIC) study.

**Estimation and Regression Analysis with Sequentially Truncated Data**

♦*Jing Qian1, Jingyao Hou1, Erik Parner2, Morten Overgaard2，Rebecca Betensky3.*

1University of Massachusetts Amherst, USA， 2Aarhus University, Denmark， 3New York University, USA

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.

**Confidence intervals for high-dimensional censored quantile regression**

*Tony SIT1and* ♦*Yu GUO1.*

1The Chinese University of Hong Kong

This work proposes a novel method for constructing confidence intervals for censored quantile regression in high-dimensional data settings where the number of covariates may significantly exceed the sample size. Building on the weighted loss function introduced by Wang (2009), we apply an L1 penalization and subsequently perform a debiasing process on the resulting estimate. The debiased estimator is shown to exhibit asymptotic normality, providing a robust basis for inference. Unlike existing research, our approach relaxes the global linearity condition to a local linearity condition near the quantile of interest, offering a more flexible and accurate model. This method is particularly advantageous when dealing with heteroskedastic effects or violations of global linearity. Simulation results demonstrate superior performance of our method in constructing confidence intervals.

**Session 24CHI80: Recent advances in Survival Data Modeling**

**Parametric and semiparametric estimation methods for survival data under a flexible class of models**

♦*Wenqing He*1*，Grace Yi* 1.

1University of Western Ontario

In survival analysis, accelerated failure time models are useful in modeling the relationship between failure times and the associated covariates, where covariate effects are assumed to appear in a linear form in the model. Such an assumption of covariate effects is, however, quite restrictive for many practical problems. To incorporate flexible nonlinear relationship between covariates and transformed failure times, we propose partially linear single index models to facilitate complex relationship between transformed failure times and covariates. We develop two inference methods which handle the unknown nonlinear function in the model from different perspectives. The first approach is weakly parametric which approximates the nonlinear function globally, whereas the second method is a semiparametric quasi-likelihood approach which focuses on picking up local features. We establish the asymptotic properties for the proposed methods. A real example is used to illustrate the usage of the proposed methods, and simulation studies are conducted to assess the performance of the proposed methods for a broad variety of situations.

**Expectation solution algorithm for the additive hazards mixture cure model with varying coefficients.**

♦*zhongwen zhang*.

Nanjing Medical University.

In survival analysis, the varying coefficient models are very effective tool to explore the dynamic pattern of covariate effect. Profit from their flexibility and interpretability, various types of varying coefficient survival models have experienced deep and exciting developments on methodological, theoretical and applied sides. However, the varying coefficients survival model still haven’t attracted enough attention when there exists a non-ignorable cure fraction in the population. In this paper, we propose an additive hazards mixture cure model with semi-varying coefficients to fill this gap. The proposed model allows some covariate effects to be time-invariant while other covariate effects to be time-varying. We develop a expectation solution algorithm based method to obtain estimating equation for the parameters in the model, and provide a fast and straightforward algorithm implementation method. The finite sample performance of the estimators is examined by simulation studies. The proposed model and estimation are illustrated with an analysis of data from UKbiobank.

**Identification of Prognostic and Predictive Subgroups for Clustered Survival Data**

♦*Ye He*1, *DongSheng Tu*2, *Liu Liu*3 *and* *Ling Zhou*4, Zhongwen Zhang

1Sichuan Normal University, 2Queen’s University, 3College of Chengdu University of Technology， 4Southwestern University of Finance and Economics.

Clinical investigators are interested in identifying of patients subgroups with heterogeneity in clinical outcomes and also patients subgroups with different treatment effects based on multiple biomarkers. These two types of subgroups are respectively termed as prognostic and predictive subgroups in the clinical literature. Existing methods can identify either prognostic or predictive subgroups separately. However, they require the outcomes independent while patients enrolled in trials we recently analyzed were from different clinical centres and, therefore, their clinical outcomes may be correlated. In this paper, we propose a novel double-centre-augmented frailty (DCAF) model to identify prognostic and predictive subgroups simultaneously based on clustered survival data. Particularly, the proposed DCAF enables simultaneous subgroup analysis and variable selection to identify prognostic subgroups as well as homogeneity fusion on high-dimensional biomarkers to identify predictive subgroups, resulting in higher estimation and clustering accuracy when there exist heterogeneity across clusters and weak signals across high-dimensional biomarkers. Instead of commonly used pairwise penalties, a differentiable centre-augmented harmonic type penalty is used to identify subgroups. The proposed DCAF was applied to the data from the companion studies of the clinical trials, which motivated this work, and resulted in respectively two prognostic and seven predictive subgroups of patients with advanced pancreatic cancer, and two prognostic and three predictive subgroups of patients with advanced colorectal cancer.

**Integrative Approach in Survival Analysis Combining Probability and Non-Probability Samples**

♦*Hua Shen*1*，Zheng Yu*1.

1University of Calgary

Survival analysis, traditionally reliant on probability sampling methods to ensure generalizability, now faces new challenges and opportunities with the popularity of large non-probability datasets, such as electronic health records and online repositories. We explore innovative methodologies for integrating probability and non-probability samples in survival analysis, aimed at enhancing the robustness and applicability of survival estimates based on mixed data sources. We demonstrate the integrative statistical method that adjusts for the biases and sampling errors that are otherwise inherent in non-probability samples.

**Session 24CHI102: Semi-Supervised and Multi-Task Learning**

**Revisiting Scalarization in Multi-Task Learning: A Theoretical Perspective**

*Yuzheng Hu*1, *Ruicheng Xian*1, *Qilong Wu*1 *Qiuling Fan*1, *Lang Yin*1, ♦Han Zhao.

1University of Illinois Urbana-Champaign

Linear scalarization, i.e., combining all loss functions by a weighted sum, has been the default choice in the literature of multi-task learning (MTL) since its inception. In recent years, there has been a surge of interest in developing Specialized Multi-Task Optimizers (SMTOs) that treat MTL as a multi-objective optimization problem. However, it remains open whether there is a fundamental advantage of SMTOs over scalarization. In fact, heated debates exist in the community comparing these two types of algorithms, mostly from an empirical perspective. In this talk, I will revisit scalarization from a theoretical perspective. I will be focusing on linear MTL models and studying whether scalarization is capable of fully exploring the Pareto front. Our findings reveal that, in contrast to recent works that claimed empirical advantages of scalarization, scalarization is inherently incapable of full exploration, especially for those Pareto optimal solutions that strike the balanced trade-offs between multiple tasks. More concretely, when the model is under-parametrized, we reveal a multi-surface structure of the feasible region and identify necessary and sufficient conditions for full exploration. This leads to the conclusion that scalarization is in general incapable of tracing out the Pareto front. Our theoretical results provide a more intuitive explanation of why scalarization fails beyond non-convexity.

**Soft phenotyping for sepsis via EHR time-aware semi-supervised soft clustering**

♦*Anru Zhang*1.

1Duke University

Sepsis is one of the most serious hospital conditions associated with high mortality. Sepsis is the result of a dysregulated immune response to infection that can lead to multiple organ dysfunction and death. Due to the wide variability in the causes of sepsis, clinical presentation, and the recovery trajectories identifying sepsis sub-phenotypes is crucial to advance our understanding of sepsis characterization, identifying targeted treatments and optimal timing of interventions, and improving prognostication. We develop a time-aware semi-supervised soft clustering algorithm guided by clinical context to identify sepsis sub-phenotypes using data from the EHR. We identified six novel sepsis hybrid sub-phenotypes and evaluated them for medical plausibility. In addition, we built an early-warning sepsis prediction model using logistic regression. Our results suggest that these novel sepsis hybrid sub-phenotypes are promising to provide more precise information on the recovery trajectory which can be important to inform management decisions and sepsis prognosis.

**How Does Semi-Supervised Learning with Pseudo-labelers Work? A Case Study**

*Yiwen Kou*1, *Zixiang Chen*1,♦*Yuan Cao*2 *and Quanquan Gu*1.

1University of California, Los Angeles 2The University of Hong Kong

Semi-supervised learning is a popular machine learning paradigm that utilizes a large amount of unlabeled data as well as a small amount of labeled data to facilitate learning tasks. While semi-supervised learning has achieved great success in training neural networks, its theoretical understanding remains largely open. In this paper, we aim to theoretically understand a semi-supervised learning approach based on pre-training and linear probing. In particular, the semi-supervised learning approach we consider first trains a two-layer neural network based on the unlabeled data with the help of pseudo-labelers. Then it linearly probes the pre-trained network on a small amount of labeled data. We prove that, under a certain toy data generation model and two-layer convolutional neural network, the semisupervised learning approach can achieve nearly zero test loss, while a neural network directly trained by supervised learning on the same amount of labeled data can only achieve constant test loss. Through this case study, we demonstrate a separation between semi-supervised learning and supervised learning in terms of test loss provided the same amount of labeled data.

**Distributed Semi-Supervised Statistical Learning**

*Jiyuan Tu*1, *Weidong Liu*2, ♦*Xiaojun Mao*2

1Shanghai University of Finance and Economics, 2Shanghai Jiao Tong University

Semi-supervised learning enhances machine learning using unlabeled data. This talk includes two parts. The first part explores two distributed semi-supervised algorithms for the generalized linear model (GLM), proving improved convergence rates with sufficient unlabeled data. This method requires fewer communication rounds than fully-supervised approaches. In the second part, we further introduce a multi-round distributed debiased estimator for semi-supervised sparse statistical inference, which boosts statistical rates and handles non-smooth loss functions efficiently with efficient computations. We demonstrate the effectiveness of our methods by presenting simulation studies and real data applications that highlight the benefits of incorporating unlabeled data.

**Session 24CHI31: Emerging Theory and Methods in Data Science**

**Estimation of Over-parameterized Models from an Auto-Modeling Perspective.**

*Yiran Jiang*1 *and* ♦*Chuanhai Liu*2.

1Department of Biostatistics, Yale University Chuanhai Liu, 2Department of Statistics, Purdue University

From a model-building perspective, we propose a paradigm shift for fitting over-parameterized models. Philosophically, the mindset is to fit models to future observations rather than to the observed sample. Technically, given an imputation method to generate future observations, we fit over-parameterized models to these future observations by optimizing an approximation of the desired expected loss function based on its sample counterpart and an adaptive duality function. The required imputation method is also developed using the same estimation technique with an adaptive m-out-of-$n$ bootstrap approach. We illustrate its applications with the many-normal-means problem, n < p linear regression, and neural network-based image classification of MNIST digits. The numerical results demonstrate its superior performance across these diverse applications. While primarily expository, the corresponding paper conducts an in-depth investigation into the theoretical aspects of the topic. It concludes with remarks on some open problems.

**Knowledge Cascade: Reverse Knowledge Distillation on Nonparametric Multivariate Functional Estimation**

♦*Luyang Fang*1, *Haoran Lu*1, *Yongkai Chen*1, *Wenxuan Zhong*1, *Ping Ma*1.

1University of Georgia

Knowledge distillation methods have achieved remarkable performance in compressing the knowledge learned by a large-and-complex model (teacher) to a small-and-simple model (student), resulting in easier deployment while maintaining performance. However, classic knowledge distillation methods highly depend on a well-trained teacher model, which can be computationally expensive to train and may not always be available. Motivated by the challenges in developing teacher models, we propose the knowledge cascade (KCas), a reversed version of knowledge distillation that utilizes the knowledge learned by the student model to help train the teacher model. Although this is challenging since the teacher model often contains more information, we show that KCas is possible by taking advantage of statistical theories. We demonstrate KCas on the nonparametric multivariate functional estimation in reproducing kernel Hilbert space. One crucial problem is the daunting computational cost of selecting smoothing parameters, whose number increases exponentially with the number of predictors. KCas transfers the knowledge of smoothing parameters learned from the student model to the teacher model based on empirical and asymptotic results, significantly reducing the computational burden in high-dimensional and large datasets. Empirical evaluations on simulated and real data demonstrate the effectiveness of our KCas method.

**WEST: An Ensemble Method for Spatial Transcriptomics Analysis**

♦*Jiazhang Cai*1, *Huimin Cheng*1, *Shushan Wu*1, *Guo-Cheng Yuan*2, *Wenxuan Zhong*1, *Ping Ma*1

1University of Georgia, 2Icahn School of Medicine at Mount Sinai

Spatial transcriptomics emerges as a groundbreaking technology, enabling simultaneous profiling of gene expression and spatial orientation within biological tissues. Yet, when analyzing spatial transcriptomics data, effective integration of expression and spatial information poses considerable analytical challenges. Although many methods have been developed to address this issue, many are platform-specific and lack the general applicability to analyze diverse datasets. In this article, we propose a novel method called the Weighted Ensemble method for Spatial Transcriptomics (WEST) that utilizes ensemble techniques to improve the performance and robustness of spatial transcriptomics data analytics. We compare the performance of WEST with five popular methods on both synthetic and real-world datasets. WEST represents a significant advance in detecting spatial domains, offering improved accuracy and flexibility compared to existing methods, making it a valuable tool for spatial transcriptomics data analytics.

**Pseudo-Labeling for Kernel Ridge Regression under Covariate Shift**

♦*Kaizheng Wang*1

1Columbia University

We develop and analyze a principled approach to kernel ridge regression under covariate shift. The goal is to learn a regression function with small mean squared error over a target distribution, based on unlabeled data from there and labeled data that may have a different feature distribution. We propose to split the labeled data into two subsets and conduct kernel ridge regression on them separately to obtain a collection of candidate models and an imputation model. We use the latter to fill the missing labels and then select the best candidate model accordingly. Our non-asymptotic excess risk bounds show that in quite general scenarios, our estimator adapts to the structure of the target distribution as well as the covariate shift. It achieves the minimax optimal error rate up to a logarithmic factor. The use of pseudo-labels in model selection does not have major negative impacts.

**Session 24CHI64: Recent Advances in Design Of Experiments and Subsampling**

**Cyclic Bayesian D-optimal augmented definitive screening designs**

♦*Chunyan Wang1, Jinyu Yang2, Min-Qian Liu2*

1Renmin University of China, 2Nankai University

Experimenters often prefer quantitative factors with three levels since they allow the assessment of curvature in the factor-response relationship, especially in the fields of chemical science and engineering. Therein, definitive screening designs (DSDs), introduced by Jones and Nachtsheim (2011), have been widely used for screening as they make it possible to identify main effects and second-order effects in one step. However, as supersaturated designs for second-order models, DSDs do not always behave well due to the severe aliasing among the second-order effects. Particularly when the number of active effects is more than half of the run size of a DSD or different analysis methods produce very inconsistent results, a follow-up experiment is needed to clarify the initial results. Incorporating the information obtained from the initial design, we consider the augmented design with a cyclic structure for reducing the aliasing among the second-order effects. Methods for constructing efficient augmented designs based on the Bayesian D-optimality criterion are proposed. The methods are easy to implement. Simulation results show that these augmented designs perform well in certain settings.

**Efficient model-free subsampling method for massive data**

*Zheng Zhou*1, *Zebin Yang*2, *Aijun Zhang*2, ♦*Yongdao Zhou*1

1Nankai University, 2The University of Hong Kong

Subsampling plays a crucial role in tackling problems associated with the storage and statistical learning of massive datasets. However, most existing subsampling methods are model-based, which means their performances can drop significantly when the underlying model is misspecified. Such an issue calls for model-free subsampling methods that are robust under diverse model specifications. Recently, several model-free subsampling methods have been developed. However, the computing time of these methods grows explosively with the sample size, making them impractical for handling massive data. In this article, an efficient model-free subsampling method is proposed, which segments the original data into some regular data blocks and obtains subsamples from each data block by the data-driven subsamplingmethod. Compared with existing model-free subsampling methods, the proposed method has a significant speed advantage and performsmore robustly for datasets with complex underlying distributions.As demonstrated in simulation experiments, the proposed method is an order ofmagnitude faster than other commonly used model-free subsampling methods when the sample size of the original dataset reaches the order of 10^7.

Moreover, simulation experiments and case studies show that the proposed method is more robust than other model-free subsampling methods under diverse model specifications and subsample sizes.

**Sequential Order-of-addition Experiments**

*Haojie Man*1 and ♦*Jianbin Chen*1.

1Beijing Institute of Technology.

The order-of-addition (OofA) experiment has received a great deal of attention in the recent literature. The primary goal of the OofA experiment is to identify the optimal order in a sequence of m components. However, existing models and methods primarily focus on predicting response values after a given sequence, lacking the capability to achieve precise global optimization. Especially in scenarios with a large number of components, current methods often incur substantial computational costs. In this article, we propose a method for identifying variable importance in OofA experiments based on GP model and the Hellinger distance. An efficient construction algorithm

for the criteria that achieves space-filling and balance properties to OofA components is proposed. Moreover, we present a two-stage efficient global optimization algorithm for sequential OofA experiments, which enhances the computational efficiency, especially for situations with a large number of components. Theoretical supports are given to illustrate the effectiveness of the proposed method. The proposed method can obtain the optimal order for large m efficiently. Numerical experiments are used to demonstrate the effectiveness of the proposed method.

**Stratification pattern enumerator and its applications**

♦*Ye Tian*1, *Hongquan Xu*2

1Beijing University of Posts and Telecommunications, 2UCLA

Space-filling designs are widely used in computer experiments. A minimum aberration-type space-filling criterion was recently proposed to rank and assess a family of space-filling designs including orthogonal array-based Latin hypercubes and strong orthogonal arrays. However, it is difficult to apply the criterion in practice because it requires intensive computation for determining the space-filling pattern, which measures the stratification properties of designs on various subregions. In this article, we propose a stratification pattern enumerator to characterise the stratification properties. The enumerator is easy to compute and can efficiently rank space-filling designs. We show that the stratification pattern enumerator is a linear combination of the space-filling pattern. Based on the connection, we develop efficient algorithms for calculating the space-filling pattern. In addition, we establish a lower bound for the stratification pattern enumerator and present construction methods for designs that achieve the lower bound using multiplication tables over Galois fields. The constructed designs have good space-filling properties in low-dimensional projections and are robust under various criteria.

**Session 24CHI117: Statistical Learning from a Modern Perspective**

**Towards Non-Asymptotic Convergence for Diffusion-Based Generative Models**

♦*Gen Li*1, *Yuting Wei*2，*Yuxin Chen*2.

1The Chinese University of Hong Kong , 2University of Pennsylvania.

Diffusion models, which convert noise into new data instances by learning to reverse a Markov diffusion process, have become a cornerstone in contemporary generative modeling. While their practical power has now been widely recognized, the theoretical underpinnings remain far from mature. In this work, we develop a suite of non-asymptotic theory towards understanding the data generation process of diffusion models in discrete time, assuming access to $\ell\_2$-accurate estimates of the (Stein) score functions. For a popular deterministic sampler (based on the probability flow ODE), we establish a convergence rate proportional to $1/T$ (with $T$ the total number of steps), improving upon past results; for another mainstream stochastic sampler (i.e., a type of the denoising diffusion probabilistic model), we derive a convergence rate proportional to $1/\sqrt{T}$, matching the state-of-the-art theory. Imposing only minimal assumptions on the target data distribution (e.g., no smoothness assumption is imposed), our results characterize how $\ell\_2$ score estimation errors affect the quality of the data generation process. In contrast to prior works, our theory is developed based on an elementary yet versatile non-asymptotic approach without resorting to toolboxes for SDEs and ODEs.

**Transfer learning for contextual multi-armed bandits**

*♦Changxiao Cai1, T. Tony Cai2, Hongzhe Li2*

1University of Michigan, 2University of Pennsylvania

Motivated by a range of applications, we study in this paper the problem of transfer learning for nonparametric contextual multi-armed bandits under the covariate shift model, where we have data collected from source bandits before the start of the target bandit learning. The minimax rate of convergence for the cumulative regret is established and a novel transfer learning algorithm that attains the minimax regret is proposed. The results quantify the contribution of the data from the source domains for learning in the target domain in the context of nonparametric contextual multi-armed bandits.

In view of the general impossibility of adaptation to unknown smoothness, we develop a data-driven algorithm that achieves near-optimal statistical guarantees (up to a logarithmic factor) while automatically adapting to the unknown parameters over a large collection of parameter spaces under an additional self-similarity assumption. A simulation study is carried out to illustrate the benefits of utilizing the data from the source domains for learning in the target domain.

**Generative Model for Distributionally Robust Optimization**

*Chen Xu*1, *Jonghyeok*  Lee1, *Xiuyuan* *Cheng*2， ♦*Yao Xie*3.

1Georgia Institute of Technology , 2Duke University, 3Georgia Institute of Technology

We present a generative-model-based framework, called 𝙵𝚕𝚘𝚠𝙳𝚁𝙾, for solving distributionally robust optimization (DRO) problems with Wasserstein uncertainty sets while aiming to find continuous worst-case distribution (also called the Least Favorable Distribution, LFD) and sample from it. The requirement for LFD to be continuous is so that the algorithm can be scalable to problems with larger sample sizes and achieve better generalization capability for the induced robust algorithms. To tackle the computationally challenging infinitely dimensional optimization problem, we leverage flow-based models and continuous-time invertible transport maps between the data distribution and the target distribution and develop a Wasserstein proximal gradient flow type algorithm. In theory, we establish the equivalence of the solution by optimal transport map to the original formulation, as well as the dual form of the problem through Wasserstein calculus and Brenier theorem. In practice, we parameterize the transport maps by a sequence of neural networks progressively trained in blocks by gradient descent. We demonstrate its usage in adversarial learning, distributionally robust hypothesis testing, and a new mechanism for data-driven distribution perturbation differential privacy, where the proposed method gives strong empirical performance on high-dimensional real data.

**Contractive diffusion models and score matching by continuous reinforcement learning**

♦*Wenpin Tang*

Columbia University

In this talk, I will link two different topics. The past decade has witnessed the success of generative modeling (e.g. GANs, VAEs,...) in creating high quality samples in a wide variety of data modalities. The first part of this talk is concerned with the recently developed diffusion models, the key idea of which is to reverse a certain stochastic dynamics. I will first take a continuous-time perspective, and examine the performance of different SDE schemes including VE (variance exploding) and VP (variance preserving). The discretization is more subtle, and our idea is to "contract" the reversed dynamics leading to possible new diffusion model designs.

In the second part, I will talk about continuous reinforcement learning. Reinforcement Learning (RL) has been successfully applied to wide-ranging domains in the past decade. Recent years have witnessed a fast growing body of research that has extended the frontiers of continuous RL such as designing model-free methods and algorithms. I will discuss the recently introduced “q-learning” and closely related policy optimization. Finally, I will highlight a natural application of continuous RL to fine-tune the score function in the diffusion models. If time permits, I will also describe some ongoing efforts towards diffusion alignment and direct preference optimization for large language models.

**Session 24CHI89: Recent Developments in Experimental Designs**

**Bayesian Optimal Designs for Linear Regression with High Predictive Efficiency in the Event of Model Uncertainty**

*Shanika Basnayake*1*,* ♦*Po Yang*1 *and Chang-Yun Lin*2

1University of Manitoba, 2National Chung Hsing University

It is well known that classical optimal designs depend on an assumed model which may not be a true model. To overcome this, Bayesian designs were introduced. For response surface experiments, the prediction of the response is an important task. Bayesian I-optimal designs minimize the average variance of prediction, thereby increasing the prediction efficiency. We introduce three new Bayesian optimality criteria for constructing optimal designs that have high prediction efficiency and less dependence on an assumed model. Optimal designs are obtained using the Bayesian I-criterion and the newly introduced criteria. The performance of the criteria is compared with that of existing optimality criteria using graphical methods and some efficiency measures.

**Design Construction and Model Selection for Small Mixture-Process Variable Experiments with High-dimensional Model Terms**

♦*Chang-Yun Lin*1*，Kashinath Chatterjee*2

1National Chung Hsing University, 2Augusta University

This paper considers the design construction and model selection for mixture- process variable experiments where the number of variables is large. For such ex- periments the generalized least squares estimates cannot be obtained and hence it will be difficult to identify the important model terms. To overcome these problems, here we employ the generalized Bayesian-D criterion to choose the optimal design and apply the Bayesian analysis method to select the best model. Two algorithms are developed to implement the proposed methods. A fish-patty experiment demonstrates how the Bayesian approach can be applied to a real experiment. Simulation studies show that the proposed method has a high power to identify important terms and well controls the type I error.

**Robust Active Learning for Nonlinear Regression**

♦*Xiaojian Xu*1

1Brock University

Active learning enables learners to strategically choose training data, aiming to optimize accuracy in estimation or prediction. The motivation for pursuing optimality lies in the possibility of acquiring information about the response function through sampling a relatively small number of responses. This is particularly relevant in scenarios where extensive knowledge of the input population is accessible, but obtaining responses proves to be costly or challenging. Within nonlinear regression contexts, this paper explores robust designs for active learning, taking into account potential model misspecifications. We investigate the analytical forms of optimal and robust sampling density using a minimax approach. The performance of these designs is evaluated through the integrated mean squared errors.

**Practical Design Considerations for Phase III Randomized Controlled Trials in Immuno-Oncology: A Real-Life Example**

♦*Zhouzhao Zhan*1, *Lina Wang*1, *Dongmei Lu*1 and *Mingxiu Hu*1.

1Akesobio Inc.

Adaptive and seamless phase II/III design has been established as one of the key clinical trial design options for randomized controlled trials (RCTs) in oncology for which efficient and rapid delivery of effective treatment to patients in desperate need is paramount. Recent methodological development provides opportunities for adaptations of target populations based on the data obtained from the first stage of a phase II/III seamless design and the ability to strongly control the family-wise error rate (FWER) for multiple hypotheses with co-primary populations. In this study, we consider the trial design for the development of a novel PD-1/VEGF bispecific antibody cancer treatment. In particular, we focus on the discussion of the correlations between the two test statistics at the interim and final analysis when trying to facilitate early identification of potential patient cohorts who may benefit from the treatment. Furthermore, we demonstrate the operating characteristic of the design using Monte-Carlo simulations along with a discussion on the practical aspects.

**Session 24CHI73: Recent Advances in Statistical Inference on Data Integration**

**Statistical Inference with Singular Propensity Score**

♦*Kosuke Morikawa1，Keisuke Yano2.*

1Osaka University，2Institute of Statistical Mathematics

In the fields of survey sampling, missing data analysis, and causal inference, researchers often have access to only a subset of the population of interest, which can lead to biased results. To correct this bias, weighting the estimating equations with the inverse of the propensity scores is a common method. However, this approach encounters challenges when propensity scores are extremely close to 0 or 1, as the inverse probabilities may cause divergence, thereby increasing the variance of the estimates. To address this issue, this talk introduces a singular propensity score that incorporates upper and lower bounds. We propose an information criterion specifically designed for selecting these bounds, based on observed data, and propose a new weighted estimator that aims to minimize mean squared error.

**Combining Probability and Non-probability Samples Using Semi-Parametric Quantile Regression and a Non-Parametric Estimator of the Participation Probability**

*Emily Berg1, Sixia Chen2，*♦*Cindy Yu1.*

1Iowa State University, 2University of Oklahoma

Non-probability samples are prevalent in various fields, such as biomedical studies, educational research, and business investigations, owing to the escalating challenges associated with declining response rates and the cost-effectiveness and convenience of utilizing such samples. However, relying on naive estimates derived from non-probability samples, without adequate adjustments, may introduce bias into study outcomes. Addressing this concern, data integration methodologies, which amalgamate information from both probability and non-probability samples, have demonstrated effectiveness in mitigating selection bias. Commonly employed data integration approaches encompass mass imputation, propensity score weighting, and hybrid methodologies. Nonetheless, the efficacy of these methods hinges upon the assumptions underlying the models. This paper introduces innovative and robust data integration approaches, notably a semi-parametric quantile regression-based mass imputation approach and a doubly robust approach that integrates a non-parametric estimator of the participation probability for non-probability samples. Our proposed methodologies exhibit greater robustness compared to existing parametric approaches, particularly concerning model misspecification and outliers. Theoretical results are established, including variance estimators for our proposed estimators. Through comprehensive simulation studies and real-world applications, our findings demonstrate the promising performance of the proposed estimators in reducing selection bias and facilitating valid statistical inference. This research contributes to the advancement of robust methodologies for handling non-probability samples, thereby enhancing the reliability and validity of research outcomes acrossdiverse domains.

**Inductive Matrix Completion Through Transfer Learning**

Xiaojun Mao1, ♦Hengfang Wang2, Zhonglei Wang3

1Shanghai Jiao Tong University, 2Fujian Normal University, 3Xiamen University

The advent of big data has facilitated the development of advanced models by enabling the storage of vast amounts of information. Transfer learning, a machine learning technique, transfers knowledge between different domains by using pretrained models from a source domain to enhance the performance in a target domain. In this talk, we explores inductive matrix completion within the transfer learning framework, with our proposed approach assuming group sparsity for the difference between the core matrices of the target and source domains. We explore the theoretical guarantees of our method to demonstrate the benefits of transfer learning over existing approaches. Several synthetic experiments are conducted to evaluate the proposed method, showing that it outperforms existing methods.

**Estimation of Treatment Effects without Ignorability Using Observational Studies**

♦*Guoliang Ma*1, *Cindy Yu*2，*Zhonglei Wang*1

1Xiamen University, 2Iowa State University

For observational studies, the assignment to treatment is often affected by auxiliary information or even the unobserved outcome of interest. Existing works for causal inference mainly assume ignorable assignment, and it may lead to erroneous inference if such an assumption is wrong. In this paper, we propose a novel semiparametric iterative procedure for causal inference based on a general non-ignorable assignment assumption, and it (a) can consistently estimate the model parameters with controllable bias, (b) allows for heterogeneity among individuals, and (c) is flexible in identifying a wide range of causal effects. Limiting properties of the proposed procedure are rigorously investigated. We conduct extensive simulation studies under various settings and demonstrate the efficacy of the proposed ES algorithm. Application to the National Longitudinal Survey of the Young Men on the 1997 cohort reveals that the 1997 cohort relied on potential outcomes to make college-entrance decisions. On average, the return on college experience is large and grows over the years.

**Session 24CHI69: Recent Advances in Modern Data Science**

**Partially-Global Fréchet Regression**

*Danielle Tucker，* ♦*Yichao Wu.*

We propose a partially-global Frechet regression model by extending the profiling technique for the partially linear regression model (Severini and Wong 1992). This extension allows for the response to come from a generic metric space and can incorporate a combination of Euclidean predictors and a predictor which comes from another generic metric space. By melding together the local and global Frechet regression models proposed by Petersen and Muller (2019), we gain a model that is more flexible than global Frechet regression and more accurate than local Frechet regression when the data generating process relies on a non-Euclidean predictor or is truly “global (linear)” for some scalar predictors. In this paper, we provide theoretical support for partially-global Frechet regression and demonstrate its competitive finite-sample performance when applied to both simulated data and to real data which is too complex for traditional statistical methods.

**MR-SPLIT: A Robust Method in One-Sample Mendelian Randomization Studies**

*Ruxin Shi1, Ling Wang1, Stephen Burgess2，*♦*Yuehua Cui1.*

1Michigan State University，2University of Cambridge

Mendelian Randomization (MR) is a widely embraced approach to assess causality in epidemiological studies. The two-stage least squares (2SLS) method is a commonly adopted approach in MR analysis with individual level data. However, it can lead to biased estimates when instrumental variables (IVs) are weak (weak instrument bias). Moreover, when IV selection and causal estimation are done with the same data, it leads to the IV selection bias or the winner’s curse problem. Focusing on one-sample MR analysis, we propose a robust method known as Mendelian randomization with adaptive sample-splitting with cross-fitting instruments (MR-SPLIT), designed to address issues related to weak instruments as well as IV selection bias. A multi-sample splitting strategy is introduced to enhance the robustness of type I error control and improve statistical power. Comprehensive simulation studies were carried out to compare the method's performance against its counterparts, showcasing its superiority in terms of bias reduction, effective type I error control, and increased power. We applied the method to a real dataset and identified biologically meaningful causal effect. Our method provides a robust solution to address IV selection and weak instrument bias issues in one-sample MR analyses.

**Tight Community Detection**

*Jiayi Deng1, Xiaodong Yang2, Jun Yu3, Jun Liu2, Zhaiming Shen4 and* ♦*Huimin Cheng5.*

1Renmin University of China，2Harvard University，3Beijing Institute of Technology, 4University of Georgia ，5Boston University

Conventional community detection methods often categorize all nodes into clusters. However, the presumed community structure of interest may only be valid for a subset of nodes (named as “tight nodes”), while the rest of the network may consist of noninformative “scattered nodes”. For example, a protein-protein network often contains proteins that do not belong to specific biological functional modules but are involved in more general processes, or act as bridges between different functional modules. Forcing each of these proteins into a single cluster introduces unwanted biases and obscures the underlying biological implication. To address this issue, we propose a tight community detection (TCD) method to identify tight communities excluding scattered nodes. The algorithm enjoys a strong theoretical guarantee of tight node identification accuracy and is scalable for large networks. The superiority of the proposed method is demonstrated by various synthetic and real experiments.

**Online Selective Conformal Inference**

*Yajie Bao1, Yuyang Huo1, Haojie Ren1,*

*♦Changliang Zou2*

1Shanghai Jiao Tong University, 2Nankai University

We study the problem of post-selection predictive inference in an online fashion. To avoid devoting resources to unimportant units, a preliminary selection of the current individual before reporting its prediction interval is common and meaningful in online predictive tasks. Since the online selection causes a temporal multiplicity in the selected prediction intervals, it is important to control the real-time false coverage-statement rate (FCR) which measures the overall miscoverage level. We develop a general framework named CAP (Calibration after Adaptive Pick) that performs an adaptive pick rule on historical data to construct a calibration set if the current individual is selected and then outputs a conformal prediction interval for the unobserved label. We provide tractable procedures for constructing the calibration set for popular online selection rules. We proved that CAP could achieve an exact selection-conditional coverage guarantee in the finite-sample and distribution-free regimes. To account for the distribution shift in online data, we also embed CAP into some recent dynamic conformal prediction algorithms and show that the proposed method can deliver long-run FCR control. Numerical results on both synthetic and real data corroborate that CAP can effectively control FCR around the target level and yield more narrowed prediction intervals over existing baselines across various settings.

**Session 24CHI66: Recent Advances in High-Dimensional Statistical Inference**

**Direct estimation and inference of higher-level correlations from lower-level data**

♦*Yue Wang1, Dan Cheng2, Haoran Shi2*

1Colorado School of Public Health, 2Arizona State University

The gut microbiome's pivotal role in host health is well-established. A growing body of evidence underscores the complex interplay between gut microbes and human metabolism, hinting at a potential metabolome-driven mechanism that underlies the connection between the host and its microbiome. Notably, many metabolites, including secondary bile acids, exhibit similar patterns in their interactions with certain gut microbes. In this presentation, I will introduce a partially linear model designed to measure the similarity between metabolites with regard to their interactions with gut microbes. Our approach offers great flexibility in accounting for important confounding variables like diet and age. Unlike marginal correlations between metabolites, our microbial correlation method unveils the intricate mechanisms governing the connections between two metabolites in the context of distinct gut microorganisms. I will also showcase an application of this methodology using data from Inflammatory Bowel Disease (IBD) patients, shedding light on the intricate interactions between metabolites in individuals affected by IBD.

**Asset Splitting Algorithm for Ultrahigh Dimensional Portfolio Selection and Its Theoretical Property**

*Zhanrui Cai1, Changcheng Li2, Jiawei Wen3，* ♦*Songshan Yang4.*

1The University of Hong Kong, 2Dalian University of Technology, 3Meta, 4Renmin University of China

The presence of a huge number of assets poses challenges to classical portfolio selection algorithms. Constrained L1 minimization approaches have been proposed to directly estimate effective parameters in the optimal portfolio. Linear programming method and alternating direction method of multiplier (ADMM) algorithm is used to solve the corresponding minimization problems. However, these two algorithms may fail due to the limitations of computing time and computing memory when a huge number of assets are considered in the portfolio optimization. This article proposes an asset splitting ADMM (AS-ADMM for short), a parallel computing algorithm, to tackle such challenges, and establishes the convergence property of the new algorithm. Furthermore, we develop a new regularization method for estimating the effective parameters with the folded-concave penalty and establish its oracle property. The local linear approximation (LLA) algorithm is used to redirect the new method to a weighted L1 regularization method. We conduct simulation studies to investigate the advantage of the proposed algorithm and regularized model in solving the high dimensional portfolio selection problems. A real data example is also included to demonstrate the applicability of the proposed algorithms and regularization methods.

**Network-Informed High-Dimensional Analysis Framework for Gene Set Discovery and Disease Risk Prediction**

*Yihao Wang1, Yue Wang2，* ♦*Jin Jin1*

1Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania，2Department of Biostatistics and Informatics, University of Colorado Anschutz Medical Campus

Tests on single genes for their marginal association with disease outcomes can be underpowered, especially in the presence of limited sample sizes, while jointly analyzing gene sets that collectively participate in specific biological functions and efficiently leveraging external information on gene interactions could lead to increased power and biologically more insightful findings. We propose a method for identifying differentially expressed gene pathways between health conditions by efficiently leveraging external gene regulatory network (GRN) information. Our simulation and real data analyses have illustrated the potential of our method in fundamentally improving the test power while having a good type I error control. Based upon the same framework, we next utilize the identified significant pathways and the corresponding network information under the linear discriminant analysis (LDA) framework to improve classification of disease status. Our analyses on a TCGA dataset for lung cancer patients, where we identified differentially expressed gene pathways and utilized the corresponding gene expression data to classify lung cancer stages with external KEGG pathway information, showed evidence of our method improving upon existing classification algorithms.

**Nonparametric Composition-on-composition Regression Analysis for High Dimensional Microbiome Data**

♦*Xiang Zhan*

Peking University

High-dimensional compositional data are frequently encountered nowadays in scientific research of many disciplines, such as microbiome research considered in this article. Regression analysis with compositional data being either responses or predictors have been well studied. However, when both responses and predictors are compositional, the inventory of analysis tools is surprisingly limited. Among the few existing methods, most of them rely on a log-ratio transformation to move compositional data analysis from simplex to reals. Yet, a serious weakness of these methods is failure to handle the substantial fraction of zeroes observed in microbiome data. To investigate associations between multiple high-dimensional microbial compositions, we propose a Nonparametric Composition-On-Composition (NCOC) regression analysis method which does not require log-ratio transformations and hence can handle zeroes in the data. To account for high dimensionality, we estimate regression coefficients using a penalized estimation equation approach to improve its accuracy. Finally, statistical inference procedures are proposed to quantify uncertainty in our model predictions. Both superior performance of NCOC along with the validity and potential usefulness of our inference procedures are demonstrated through comprehensive numerical simulations studies and real data application and case studies.

**Session 24CHI65: Recent Advances in High Dimensional Complex Data Analysis**

**The Identifiability of Copula Models of Competing Risks Data with Exponentially Distributed Margins**

♦*Antai Wang1.*

1New Jersey Institute of Technology

We present the identi ability property of Archimedean copula models for dependent competing risks data when at least one of the failure times is ex-ponentially distributed. With this property, it becomes possible to quantify the dependence between competing events based on exponentially distributed ependent censored data. We demonstrate our estimation procedure using

simulation studies and in an application to survival data.

**A Tree-Based Conformal Prediction Method for Multi-Label Classification**

*Chhavi Tyagi1，* ♦*Wenge Guo1.*

1New Jersey Institute of Technology

Multi-label classification is a common challenge in various machine learning applications, where a single data instance can be associated with multiple labels simultaneously. In this paper, we introduce a novel tree-based method for multi-label classification using conformal prediction and multiple testing tools. The proposed method employs hierarchical clustering with labelsets to develop a hierarchical tree, which is then formulated as a multiple testing problem with a hierarchical structure. The split-conformal prediction method is used to obtain marginal conformal p-values for each tested hypothesis, and two hierarchical testing procedures are developed based on marginal conformal p-values, including a hierarchical Bonferroni procedure and its modification for controlling the family-wise error rate. The prediction sets are thus formed based on the testing outcomes of these two procedures. We establish a theoretical guarantee of valid coverage for the prediction sets through proven family-wise error rate control of those two procedures. We demonstrate the effectiveness of our method in a simulation study and two real data analysis compared to other conformal methods for multi-label classification.

**Npclust: A New Horizon in Visualizing and Testing Clustered Data Effect Sizes**

♦*Yue Cui1，Solomon Harrar2*

1Missouri State University， 2University of Kentucky

In pre-post factorial designs that involve clustered units, traditional analysis often leans on parametric methods such as generalized linear mixed effects models to account for correlations within subjects. Yet, the stringent distributional assumptions of these parametric models may not always hold or be easily verified, particularly with non-metric data scales that are prevalent in outcomes like quality-of-life measures. Moreover, the realm of incomplete clustered data with missing values has been somewhat underserved by the lack ofcomprehensive tools within R for applying recent methodological advancements. To address this gap, 'npclust' has been developed, offering nonparametric effect size-based teststailored for clustered data within factorial designs that featurepre-post measurements. This approach enables hypothesis testing akin to conventional parametric tests on group means, providing interpretable effect size comparisons both over time and across various intervention groups. In addition, 'npclust' furnishes methods for multiple comparisons and the computation of simultaneous confidence intervals for these estimated effects, complemented by straightforward visualization options. The utility of 'npclust' is demonstrated through case studies in medical and epidemiological research, showcasing its applicability and interpretive power.

**Session 24CHI60: Recent Advances and Challenges in Complex Statistical Modeling and Data Analysis**

**A Model-Free Multivariate Non-Recursive Feature Elimination for Feature Selection on High-Dimensional Complex Mltiple Response Data**

♦*Siwei Xia1.*

1College of Mathematics and Physics, Chengdu University of Technology

For high-dimensional complex multiple response problems, the inherent complexity of predictors and responses is often difficult to anticipate in advance. This paper proposes a model-free multivariate feature selection technique that does not rely on a specific model to address the challenges. The proposed method effectively handels various types of multiple response datasets, such as continuous and categorical data, and allows for sparsity in both the responses and predictors. We prove that the proposed method achieves selection consistency under mild regularity conditions. We validate the performance of our method through simulations on various datasets, including those generated by generalized linear models and binary classification models with multiple responses. Additionally, we apply our method to a real-world example to demonstrate its superiority.

**Non-Zero Block Selector: A Linear Correlation Coefficient Measure for Blocking-Selection Models**

*Weixiong Liang1，* ♦*Yuehan Yang1.*

1Central University of Finance and Economics

Multiple-group data is widely used in genomic studies, finance, and social science. This study investigates a block structure that consists of covariate and response groups. It examines the block-selection problem of high-dimensional models with group structures for both responses and covariates, where both the number of blocks and the dimension within each block are allowed to grow larger than the sample size. We propose a novel strategy for detecting the block structure, which includes the block-selection model and a non-zero block selector (NBS). We establish the uniform consistency of the NBS and propose three estimators based on the NBS to enhance modeling efficiency. We prove that the estimators achieve the oracle solution and show that they are consistent, jointly asymptotically normal, and efficient in modeling extremely high-dimensional data. Simulations generate complex data settings and demonstrate the superiority of the proposed method. A gene-data analysis also demonstrates its effectiveness.

**Ensemble Methods for Testing A Global Null**

♦*Yaowu Liu*, *Zhonghua Liu*, *Xihong Lin*

Testing a global null is a canonical problem in statistics and has a wide range of applications. In view of the fact that no uniformly most powerful test exists, prior and/or domain knowledge are commonly used to focus on a certain class of alternatives to improve the testing power. However, it is generally challenging to develop tests that are particularly powerful against a certain class of alternatives. In this paper, motivated by the success of ensemble learning methods for prediction or classification, we propose an ensemble framework for testing that mimics the spirit of random forests to deal with the challenges. Our ensemble testing framework aggregates a collection of weak base tests to form a final ensemble test that maintains strong and robust power. We apply the framework to four problems about global testing in different classes of alternatives arising from Whole Genome Sequencing (WGS) association studies. Specific ensemble tests are proposed for each of these problems, and their theoretical optimality is established in terms of Bahadur efficiency. Extensive simulations and an analysis of a real WGS dataset are conducted to demonstrate the type I error control and/or power gain of the proposed ensemble tests.

**Maximum Entropy Estimator for Hidden Markov Models: Reduction to Dimension 2**

♦*Shulan Hu*

Approximate message passing (AMP) emerges as an effective iterative algorithm for solving high-dimensional statistical problems. However, prior AMP theory, which focused mostly on high-dimensional asymptotics, fell short of predicting the AMP dynamics when the number of iterations surpasses o(log n / log log n) (with n the problem dimension). To address this inadequacy, this talk introduces a non-asymptotic framework towards understanding AMP. Built upon a new decomposition of AMP updates in conjunction with well-controlled residual terms, we lay out an analysis recipe to characterize the finite-sample convergence of AMP up to O(n / polylog(n)) iterations. We will discuss concrete consequences of the proposed analysis recipe in the Z2 synchronization problem; more specifically, we predict the behavior of randomly initialized AMP for up to O(n/poly(\log n)) iterations, showing that the algorithm succeeds without the need of a careful spectral initialization and also a subsequent refinement stage (as conjectured recently by Celentano et al.)

**Session 24CHI54: New Technology and Quantitative Tools in Drug Development**

**Modeling Time-to-Event Data with Brms Package**

♦*Cong Zhang1, Sebastian Weber1 and Matt Whiley1.*

1Novartis

The R package brms (Bayesian regression models using Stan) provides a powerful and flexible modeling framework, which is essentially an easy to use frontend for the Stan statistical modeling platform. Here we provide a case study to demonstrates:

•modeling censored time-to-event data in brms with borrowing fromhistorical control data,

•the use of custom contrasts allowing for custom definition of themodel parametrization, which inturn allows to specify priors on relevant bits of the problem (a priori we expect the treatment effect not to differ substantially between standard of cares),

•an additional alternative approaches of using custom coded contrasts to include with greater flexibility historical data.

**Advanced Adaptive Confirmatory Clinical Trial Designs with R**

♦*Xin Zhang1.*

1Pfizer Inc.

In drug and vaccine development, it is critical to use cutting-edge tools for designing clinical trials, which allows us the flexibility to explore a wide range of design options along with efficient use of time and resources. The rpact (R Programming for Adaptive Clinical Trials) is a modern open-source R package that provides state-of-the-art tools for clinical trial planning, evaluation, simulation and data analysis. The package is fully validated, well documented, and compliant with FDA/GxP guidelines. In this talk, we will first provide a step-by-step illustration of how to use rpact to develop adaptive designs for sample size re-calculation and population enrichment. We will also cover some basic theories behand adaptive designs, which are necessary for properly use of rpact. Next, we will provide an example of using rpact for extensive evaluation of an adaptive enrichment design for an internal project. Finally, we will share our experiences to promote the use of rpact within our statistics organization.

**Statistical Considerations for One Subgroup and Overall Population Being the Primary Analysis Populations**

*Yiqi Zhao1, Ping Yan2 ，* ♦*Xinfeng Yang3*

1Guangzhou Culture and Tourism Industry Promotion Center, Guangzhou Tourism Information and Assistance Service Center， 2Department of Data Science, Shanghai Junshi Biosciences Co., Ltd.， 3Department of Biometrics, Hengrui Pharmaceuticals, Inc.

We will focus on how to properly design clinical trials when one subgroup and overall population are the primary analysis populations. We will show that the parameters of the overall population and the positive subgroup need to satisfy certain constraints, otherwise the parameters may lead to contradictions, and aslo show one method (called combination method) is helpful in determining α-splitting strategy between the primary endpoints. Besides, we demonstrate through three real-word applications that the combination method is helpful in the planning stage of clinical trials: guiding parameter setting in sample-size calculation assumptions, determining which subpopulation to be enrolled properly, and more accurately estimating number of events needed to reach a certain power.

**Characteristics of Regional Consistency in Multi-Regional Clinical Trials With Group Sequential Designs**

♦*Yanxin Jiang*, *Xiaoting Gao*, *Qingjia Ke* and *Man Zhao*.

In recent years, there has been a growing trend towards utilizing Multi-Regional Clinical Trials (MRCTs) for simultaneous global development of new drugs. This development strategy is based on the ICH E17 guideline. One of the keyassumptions for this strategy is that there is consistency of treatment effects across the different regions. This regional consistency can be established based on the totality of data, including the data from the early to late phases. In this holistic approach, data from the confirmatory MRCTs in a development program typically play a major role in establishing regional consistency. The capability of an MRCT to support consistency evaluation is also commonly a regulatory request or requirement by a health authority such as the China Center of Drug Evaluation. Accordingly, an MRCT is often designed with an adequate probability, e.g. 80%, for establishing the consistency in efficacy results between a country/region such as China and the entire MRCT. Quantitatively, the probability of regional consistency is often calculated using the so called Japan Methods 1 and 2. In practice, a complexity in assessing probability of consistency arises when an MRCT has interim analyses based on which the trial could be stopped early for efficacy. The question is then whether the confirmatory MRCT would have an adequate probability of establishing regional consistency if the trial would be stopped early. In addition, there is also a keen interest in determining the overall probability of achieving regional consistency in a clinical trial with interim analysis, which could be stopped early for efficacy. As a relatively common type of statistical designs for MRCTs, a group sequential design (GSD) is utilized with the option of stopping the trial for efficacy based on an interim analysis. For this class of practically useful scenarios, we conducted investigations of the operating characteristics of the probability of consistency. This talk will present the results of our investigations and the useful insight we have obtained for guiding the designs of future MRCTs using GSDs where the trials could be stopped early for efficacy.

**Session 24CHI47: New Development in Causal Discovery and Biomarker Identification for Precision Medicine**

**Causal Inference Meets Non-Sparse High-Dimensional Confounders: A Remedy Using Feature Aggregation and Data Integration**

♦*Jinyuan Liu1.*

1Vanderbilt University

Inference for causal effects in the presence of high-dimensional confounders for observational studies is challenging. Recent development focuses on data-adaptive approaches such as machine learning (ML) with regularizations to handle high-dimensional confounders, especially when p>>n. However, the induced regularization bias and overfitting may cause substantial bias in naive estimators of causal effects. Albeit the advancement of debiased ML or target learning that employs both orthogonal scores and cross-fitting to resolve such challenges in part, one is still subject to omitted variable bias (OVB) when the sparsity assumption imposed on the confounders is violated. Moreover, such conditions become prevalent, incentivized by technological innovations such as high-throughput sequencing and wearable devices. Here, we propose a remedy using feature aggregation to bypass regularization bias. By first integrating the confounding information into non-linear distance matrices, we deploy the efficient influence function (EIF) for between-subject metrics to estimate the preserved causal parameters. Hence, the resulting doubly robust estimators enjoy superior statistical power and enhanced computational efficiency by avoiding cross-fitting or hyper-parameter tunning that causes excessive variability. This approach is exemplified by the timely issue of the “gut-brain axis” among the human microbiome, cognitive, and psychosocial measurements.

**A General Framework for Incorporating Identification Uncertainty in Individualized Treatment Rules**

♦Muxuan Liang1, Ting Ye2，Ying-Qi Zhao3.

1University of Florida， 2University of Washington， 3Fred Hutchinson Cancer Center

Estimating individualized treatment rules (ITRs) from observational data or clinical trials with non-adherence is challenging due to possible unmeasured confounding bias. Partial identification approaches using an instrumental variable (IV) provide characterizations on possible values of the conditional average treatment effects (CATEs). In this work, we develop a new class of `optimal' ITRs to guide treatment decisions when the CATEs are only partially identified. We define a novel value function allowing a reject option in treatment decisions under partial identification, and use that value function to define a class of IV-optimal ITRs with a reject option. The reject option informs who are susceptible to identification uncertainty and allows the use of alternative ITRs derived from other studies or outcomes for these patients. In addition, our framework allows users to control the size of subgroups receiving the reject option, taking into account the risks associated with unreliable or delayed treatment assignments. To estimate the IV-optimal ITRs with a reject option, we develop a weighted classification framework with a modified hinge loss function, where the weights are non-smooth transformations of nuisance parameters. We further propose an empirical augmented risk minimization approach that achieves a fast convergence rate even if the nuisance parameters are estimated using nonparametric or machine learning methods. Simulations and real data analysis are conducted to demonstrate the superiority of the developed framework and estimation procedure.

**A Bayesian Framework for Detecting Circadian Biomarkers in Transcriptomic Data**

Lingsong Meng1, Haocheng Ding1, Yutao Zhang1， ♦Zhiguang Huo1.

1University of Florida

Circadian rhythms, our body's intrinsic 24-hour cycles, play a pivotal role in regulating behavior, physiology, and overall well-being. Present in almost every cell in the human body, these rhythms govern numerous physiological processes, from sleep-wake cycles and body temperature fluctuations to melatonin secretion. The 2017 Nobel Prize was awarded to three researchers for their "discoveries of molecular mechanisms controlling the circadian rhythms," which has amplified the study of circadian rhythms through omics data. Consequently, the literature now boasts extensive circadian omics studies in data repositories, offering deep insights into potential circadian biomarkers. Such discoveries not only shed light on the molecular mechanisms of the clock in aging but also guide future endeavors in circadian biomarker detection. By leveraging this prior knowledge, there is potential that the statistical power in circadian biomarker detection can be substantially enhanced. While numerous algorithms for detecting circadian rhythmicity exist, most follow a frequentist approach, leaving a gap in methodologies that effectively integrate prior knowledge. To bridge this gap, we introduce a novel Bayesian framework tailored for circadian analysis. Bayesian methods, with their ability to incorporate prior knowledge, allow for a more in-depth interpretation of data, incorporating established scientific insights with new findings. To facilitate the selection of circadian biomarkers from high-dimensional transcriptomic data, we leverage the reversible jump Markov Chain Monte Carlo (rjMCMC) technique, a specialized MCMC method that addresses variable dimension challenges. The merits of our Bayesian framework for circadian biomarker detection include: i) Efficient computation, achieved through the utilization of conjugate priors, thereby facilitating Gibbs sampling since all posteriors are analytically determinable. ii) The capability to obtain parameter distributions for circadian genes, enhancing uncertainty estimation for circadian parameters. iii) Effective control of multiple testing through the false discovery rate, ensuring stringent statistical criteria for circadian biomarker detection.

**DEMO: Dose Exploration, Monitoring, and Optimization Using a Biological and Clinical Outcomes**

♦Ruitao Lin1.

1The University of Texas MD Anderson Cancer Center.

Phase 1-2 designs provide a methodological advance over phase 1 designs for dose finding by using both clinical response and toxicity. A phase 1-2 trial still may fail to select a truly optimal dose. because early response is not a perfect surrogate for long term therapeutic success. To address this problem, a generalized phase 1-2 design first uses a phase 1-2 design's components to identify a set of candidate doses, adaptively randomizes patients among the candidates, and after longer follow up selects a dose to maximize long-term success rate. In this paper, we extend this paradigm by proposing a design that exploits an early treatment-related, real-valued biological outcome, such as pharmacodynamic activity or an immunological effect, that may act as a mediator between dose and clinical outcomes, including tumor response, toxicity, and survival time. We assume multivariate dose-outcome models that include effects appearing in causal pathways from dose to the clinical outcomes. Bayesian model selection is used to identify and eliminate biologically inactive doses. At the end of the trial, a therapeutically optimal dose is chosen from the set of doses that are acceptably safe, clinically effective, and biologically active to maximize restricted mean survival time. Results of a simulation study show that the proposed design may provide substantial improvements over designs that ignore the biological variable.

**Session 24CHI51: New Methods in Nonparametric Statistics with Applications to Time Series**

**Oracle-efficient M-estimation for Single-Index Models with a Smooth Simultaneous Confidence Band**

♦Li Cai1, Lei Jin2, Jiuzhou Miao1，Suojin Wang2.

1Zhejiang Gongshang University, 2Texas A&M University

Single-index models are important and popular semiparametric models, as they can handle the problem of the “curse of dimensionality” and enjoy the flexibility of nonparametric modeling and the interpretability of parametric modeling. Most existing methods for single-index models are sensitive to outliers or heavy tailed distributions because they use the least squares criterion. An oracle-efficient robust M-estimator is proposed for single-index models and a smooth simultaneous confidence band is constructed by treating the index coefficients as nuisance parameters. Under general assumptions it is shown that the M-estimator for the nonparametric link function, based on any root n consistent coefficient index parameter estimators, is oracle-efficient. This means that it is uniformly as efficient as the infeasible one obtained by M-regression using the true single-index coefficient parameters. As a result, the asymptotic distribution of the maximal deviation between the M-type kernel estimator and the true link function is derived, and an asymptotically accurate robust simultaneous confidence band is established as a global inference tool for the link function. The proposed method generalizes the desirable uniform convergence property of ordinary least squares to the robust M-estimation, thereby effectively reducing the influence of the outliers and overcoming the sensitivity of least squares estimation induced by non-normal errors. On the other hand, it is a general approach that allows any

root n consistent coefficient parameter estimators to be applied in the procedure to make global inferences for the link function. Simulation studies with commonly encountered sample sizes are reported to support the theoretical findings. To illustrate, the proposed method is applied to the analysis of a car purchasing dataset.

**Jackknife Empirical Likelihood for The Correlation Coefficient with Additive Distortion Measurement Errors**

*Da Chen*, ♦*Linlin Dai*, *Yichuan Zhao*

The calculation of correlation coefficient can be inaccurate with the existence of confounding variables. Such confounding variables could act in an additive or multiplicative fashion. To study the additive model, previous research has shown residual-based estimation of correlation coefficients. The powerful tool of empirical likelihood has been used to construct the confidence interval for the correlation

 coefficient. However, the methods so far only perform well when sample sizes are large. With small sample size situations, the coverage of EL, for instance, can be below 90% at significant level 5%. On the basis of previous research, we propose new methods of interval estimation for the correlation coefficient using jackknife empirical likelihood, mean jackknife empirical likelihood and adjusted jackknife empirical likelihood. For better performance with small sample sizes, we also propose mean adjusted empirical likelihood. The simulation results show the best performance with mean adjusted jackknife empirical likelihood when the sample sizes are as small as 25. Real data analyses are used to illustrate the proposed approach.

**Testing Serial Dependence or Cross Dependence for Time Series with Underreporting**

*Keyao Wei,* *Lengyang Wang*， ♦*Yingcun Xia*

In practice, it is common for the data collected to be underreported, which is particularly prevalent in fields such as social sciences, ecology and epidemiology. Drawing inferences from such data using conventional statistical methods can lead to incorrect conclusions. In this paper, we study tests for serial or cross dependence in time series data that are subject to underreporting. We introduce new test statistics, develop corresponding group-of-blocks bootstrap techniques, and establish their consistency. The methods are shown to be efficient by simulation and are used to identify key factors responsible for the spread of dengue fever and the occurrence of cardiovascular disease.

**Tests for Detecting Change in Mean Vector Functions of Multivariate Functional Data with Repeated Observations**

*Wei Lin*1, ♦*Zhiping Qiu*1, *Xiaming Tu*2

1Fujian Normal University, 2Huaqiao University

In many scientific and technological fields, multivariate functional data is often repeatedly observed under different conditions within a given time. It is important to determine if the mean vector function for this type of data is consistently equal throughout the entire period. This paper introduces four novel global testing statistics that employ integrating and maximizing techniques to address this issue. The asymptotic expressions of the proposed testing statistics under the null hypothesis are derived, and their root-$n$ consistencies are established. Simulation studies are carried out to demonstrate the numerical performance of the proposed tests. The testing procedures are also exemplified by analyzing air pollution data from China.

**Session 24CHI127: Recent Advances in Analysis of Complex Networks Data**

**A Framework for Statistical Inference Via Randomized Algorithms**

♦*Zhixiang Zhang1 , Edgar Dobriban2*，*Sokbae Lee3 .*

1University of Macau， 2University of Pennsylvania， 3Columbia University

Randomized algorithms, such as randomized sketching or projections, are a promising approach to ease the computational burden in analyzing large datasets. However, randomized algorithms also produce non-deterministic outputs, leading to the problem of evaluating their accuracy. In this paper, we develop a statistical inference framework for quantifying the uncertainty of the outputs of randomized algorithms. We develop appropriate statistical methods— sub-randomization, multi-run plug-in and multi-run aggregation inference—by using multiple runs of the same randomized algorithm, or by estimating the unknown parameters of the limiting distribution. As an example, we develop methods for statistical inference for least squares parameters via random sketching using matrices with i.i.d. entries, or uniform partial orthogonal matrices. For this, we characterize the limiting distribution of estimators obtained via sketch-and-solve as well as partial sketching methods. The analysis of i.i.d. sketches uses a trigonometric interpolation argument to establish a differential equation for the limiting expected characteristic function and find the dependence on the kurtosis of the entries of the sketching matrix. The results are supported via a broad range of simulations.

**Community-Finding Methods Based on the Gaps of Dynamic Interactions for the Complex Network Data**

♦*Xiaobing Zhao1, Yufeng Xia1， Yang Kuo Li1.*

1Zhejiang University of Finance and Economics

Community clustering is the main task in the analysis of network data, and the mixed membership stochastic block model has attracted a great deal of attention in statistical clustering due to its ability to detect the overlapping communities. In this paper we propose an additive hazards function model to fit the recurrent instantaneous interactions with continuous observations, and the recurrent gap times are applied to

provide an interesting alternative for the community-finding in dynamic networks. A variational expectation-maximization algorithm is applied to obtain the estimators ofthe connectivity parameters and the latent variables. The asymptotic properties of the estimates are discussed as well. Some simulation studies and applications on the primary school temporal network data are presented to illustrate the performance of the proposed models and methodology.

**Hypothesis Testing for General Network Models**

*Kang Fu1,* ♦*Jianwei Hu1, Seydou Keita1*

1Central China Normal University

The network data has attracted considerable attention in modern statistics. In research on complex network data, one key issue is ﬁnding its underlying connection structure given a network sample. The methods that have been proposed in literature usually assume that the underlying structure is a known model. In practice, however, the true model is usually unknown, and network learning procedures based on these methods may suﬀer from model misspeciﬁcation. To handle this issue, based on the random matrix theory, we ﬁrst give a spectral property of the normalized adjacency matrix under a mild condition. Further, we establish a general goodness-of-ﬁt test procedure for the unweight and undirected network. We prove that the null distribution of the proposed statistic converges in distribution to the standard normal distribution. Theoretically, this testing procedure is suitable for nearly all popular network models, such as stochastic block models, and latent space models. Further, we apply the proposed method to the degree-corrected mixed membership model and give a sequential estimator of the number of communities. Both simulation studies and real-world data examples indicate that the proposed method works well.

**The sparsity learning of decentralized networks with expectile regression**

♦*Yu Zhang*

School of Statistics and Mathematics, Zhongnan University of Economics and Law

The growing interest in decentralized distributed sparse learning arises from its capability to tackle the computational challenges in large-scale data, without being limited by the central master node. Expectile regression is widely applied to deal with heavy-tailed noise and data heterogeneity due to its computational advantages. To address the issue of data heterogeneity in decentralized distributed sparse learning, this paper proposes a decentralized Expectile Regression method (deER) that combines the expectile regression loss with L1 penalty. We first establishes the linear convergence rate of the proposed algorithm and demonstrates the statistical properties of the method, showing that the proposed estimates achieve a near-oracle convergence rate and provide theoretical results on sparse support set recovery. The effectiveness of the proposed method is validated through numerical simulations and real data analysis.

**Session 24CHI17: AI in Bioinformatics and Biomedical Applications**

**Indentify Spatial Domains Using Cell-Cell Communication Information by Graph Convolutional Network**

*Huijin Fu*, ♦*Kui Wang*.

In our pursuit of understanding spatial gene expression variation, we introduce Cyto-CommGCN, a graph convolutional network (GCN) approach. This innovative method leverages cell-cell communication information in the analysis of spatial transcriptomics (SRT) data. By integrating graph convolution techniques, Cyto-CommGCN effectively aggregates gene expression and cell-cell communication data from neighboring cells, providing a powerful tool for identifying spatial domains with consistent expression and histological features. Our comprehensive analysis of SRT datasets using Cyto-CommGCN demonstrates its exceptional ability to detect genes with significantly enriched spatial expression patterns. Furthermore, our approach excels in revealing specific expression patterns in ligand-receptor communication, surpassing the capabilities of existing methods. In summary, Cyto-CommGCN emerges as a robust, user-friendly, and computationally efficient solution. It offers profound biological insights into the intricate spatial organization of tissue, significantly advancing our understanding of complex microenvironments.

**Prediction on the Functions of Intrinsically Disordered Proteins**

♦*Zhenling Peng*

Shandong University

Intrinsically disordered proteins and regions (IDPs and IDRs) lack stable 3D structure under physiological conditions in-vitro, are common in eukaryotes, and play a variety of crucial functions in cells. One of key features of IDRs is facilitation of protein-protein and protein-nucleic acids interactions. These disordered binding regions include molecular recognition features (MoRFs), short linear motifs (SLiMs) and longer binding domains. Recently, a new and broad class of disordered binding regions, linear interacting peptides (LlPs), was introduced and applied in the MobiDB resource. LIPs are segments in protein sequences that undergo disorder-to-order transition upon binding to a protein or a nucleic acid. They cover MoRFs, SLiMs and disordered protein-binding domains. Meanwhile, computational studies support that the IDRs serving as flexible linkers (DFLs) are abundant in nature. They are also discovered to be involved in allosteric regulation via connecting protein domains and structural elements within domains. Motivated by these observations, we developed a series of computational methods including DisoRDPbind, APOD and CLIP, which are specific to predict the general disordered binding, the functional module of DFLs and LIPs, respectively. According to the Critical Assessment of Intrinsic protein Disorder (CAID) experiment, our approaches provide competitive accuracy on the disorder function prediction. The CAID experiment is a world-wide blind test, which was established in 2018 to track the progress of the state of the art predictors of IDRs and their functions.

**A Quadruplet Unified Framework Enables Analysis of Single Cell Data**

♦*Xiangjie Li*

Changping Laboratory

Dropout events in single-cell RNA sequencing (scRNA-seq) significantly impact clustering analysis, differential expression analysis, and other downstream analyses, making the denoising of scRNA-seq data a challenging problem in current research. Most existing algorithms focus solely on imputation and denoising in the context of a single batch, neglecting the simultaneous handling of batch effects and denoising in the original gene space. To address this issue, we developed QUEST, a unified algorithm framework implemented through an autoencoder model. QUEST leverages deep metric learning principles by optimizing both quadruplet loss and reconstruction loss. This approach is versatile, supporting tasks such as visualization, clustering, and denoising in both single and multiple batch scenarios. Evaluations on various simulated and real datasets demonstrate that QUEST enhances 2-dimensional visualization and clustering performance while effectively denoising scRNA-seq data in the original gene space.

**Pre-Symptomatic Detection and Alarming of COVID Using Smartwatch Data**

♦Meng Wang

We used consumer smartwatches for the pre-symptomatic detection of coronavirus disease 2019 (COVID-19) in the work of Mishra, Wang, et.al, Nat. Biomed. Eng (2020). Both offline and online anomaly detection algorithms were developed for detecting physiological alterations due to the COVID infection. This work demonstrated the potential ability that activity tracking and health monitoring via consumer wearable devices can be used for the large-scale, real-time detection of respiratory infections, often pre-symptomatically. We further developed a real-time smartwatch-based alerting system for detecting the COVID-19 and other stress events in the work of Alavi, Bogu, Wang, et al., Nat. Med (2022). In this work, the alarming algorithms were developed to detect aberrant physiological and activity signals (heart rates and steps) associated with the onset of early infection. Our work showed that a real-time alerting system can be used for early detection of infection and other stressors and employed on an open-source platform that is scalable to millions of users.

**Computing the Role of Splicing Dysregulation by SF3B1-SUGP1-DHX15 Axis in Human Cancer**

*Jian Zhang*, *Jindou Xie*, *Ji Huang*, *Xiangyang Liu*, *Ruihong Xu*, *Wojciech Galej*, *Liang Tong*, *James Manley*, ♦*Zhaoqi Liu*.

Aberrations in mRNA splicing contribute to neoplastic transformation, tumor progression and therapeutic resistance. Although mutations in the splicing factor SF3B1 are frequent in cancers, their functional effects and regulatory mechanism are poorly understood. Here, we identify that SF3B1 mutations in CLL alter splicing of a specific subunit of the PP2A to confer post-translational MYC and BCL2 activation. In addition, SF3B1 mutations in breast cancer induce a recurrent pattern of aberrant splicing leading to activation of AKT and NF-kB, enhanced cell migration, and accelerated tumorigenesis. As well, we found MAP3K7 mis-splicing leads to the anemia characteristic of SF3B1-mutated MDS by affecting terminalerythroid differentiation. Moreover, we revealed a critical role for the spliceosomal protein SUGP1 in wild-type SF3B1 splicing. Pan-cancer analyses identified SUGP1 mutations induced use of cryptic 3'ss similar to mutant SF3B1 aberrant splicing.Finally, structural modeling of a trimeric protein complex reveals that the SUGP1-SF3B1 interaction “loops out” the SUGP1 G-patch domain, facilitating its activating interaction with another helicase DHX15. Our study thus provides an unprecedented molecular view of a protein complex essential for accurate 3’ splicing, and reveals that numerous cancer-associated mutations all disrupt the critical SUGP1-SF3B1 interaction.

**Session 24CHI35: High Dimensional Hypothesis Testing Problems**

**CAST: A Covariate-Assisted Multiple Testing Framework with Sample Splitting**

*Yixin Han*1, *Changliang Zou*1, *Lilun Du*2, *Zhaojun Wang*1, ♦Peter Chien

1Nankai University, 2City University of Hong Kong

Auxiliary information often provides significant value in economics and finance for funds selection. In this paper, we develop a unified and effective framework, called Covariate-Assisted sample Splitting Testing (CAST), which can integrate various types of side information in large-scale inference for power enhancement. Our innovative CAST first aggregates the side information via variance reduction and sample splitting techniques, and thus constructs a sequence of marginal symmetry summary statistics for data-driven threshold. The principal contribution is the enhancement of statistical power and interpretability of side information while controlling the false discovery rate (FDR). Our method can flexibly accommodate various auxiliary structures and does not involve any unknown population-related quantities, further highlighting its practicality and robustness. %Our procedure is highly flexible to implement and more powerful compared with several competing methods. both theoretically and empirically. Extensive simulation experiments and an empirical application for skilled funds selection in the U.S. financial market demonstrate the satisfactory FDR control and superior statistical power of the proposed CAST over several competing methods.

**Computationally Efficient and Data-Adaptive Change-Point Inference in High Dimension**

*Guanghui Wang*1, ♦*Long Feng*2

1East China Normal University, 2Nankai University

High-dimensional change-point inference that adapts to various change patterns has received much attention recently. We propose a simple, fast yet effective approach for adaptive change-point testing. The key observation is that two statistics based on aggregating cumulative sum statistics over all dimensions and possible change-points by taking theirmaximum and summation, respectively, are asymptotically independent under some mild conditions. Hence we are able to form a new test by combining the p-values of the maximum- and summation-type statistics according to their limit null distributions. To this end, we develop new tools and techniques to establish asymptotic distribution of the maximum-type statistic under a more relaxed condition on component wise correlations among all variables than that in existing literature. The proposed method is simple to use and computationally efficient. It is adaptive to different sparsity levels of change signals, and is comparable to or even outperforms existing approaches as revealed by our numerical studies.

**Statistical Inference for Ultrahigh Dimensional Location Parameter Based on Spatial Median**

*Guanghui Chen*1, ♦*Liuhua Peng*2, *Changliang Zou*3

1Guangzhou University, 2University of Melbourne, 3Nankai University

Motivated by the widely used geometric median-of-means estimator in machine learning, this paper studies statistical inference for ultrahigh dimensionality location parameter based on the sample spatial median under a general multivariate model, including simultaneous confidence intervals construction, global tests, and multiple testing with false discovery rate control. To achieve these goals, we derive a novel Bahadur representation of the sample spatial median with a maximum-norm bound on the remainder term, and establish Gaussian approximation for the sample spatial median over the class of hyperrectangles. In addition, a multiplier bootstrap algorithm is proposed to approximate the distribution of the sample spatial median. The approximations are valid when the dimension diverges at an exponentially rate of the sample size, which facilitates the application of the spatial median in the ultrahigh dimensional region. The proposed approaches are further illustrated by simulations and analysis of a genomic dataset from a microarray study.

**Algorithm-Agnostic Inference After Changepoint Detection**

*Yinxu Jia*1, *Jixuan Liu*1, ♦*Guanghui Wang*2, *Changliang Zou*1

1Nankai University, 2East China Normal University

To evaluate the validity of changepoints detected by certain algorithms, an apparent approach is to conduct a two-sample test on data segments surrounding the identified changepoint. However, this method often yields invalid p-values due to the "double dipping" effect, where the same data is reused for both change detection and subsequent testing, thereby compromising statistical validity. While recent methodologies have incorporated selective inference tools specifically designed for sequences ofunivariate normal means, their applicability remains limited. This paper introduces a novel framework for post change detection inference, offering broader applicability across various changepoint models and detection algorithms.

**Session 24CHI126:** **Strategic Approaches and Methodological Innovations in Early Oncology Trial Designs**

**A Model-Based Design with A Randomization Scheme Considering Pharmacokinetics Exposure for Dose Optimization in Oncology**

♦*Jun Zhang.*

The primary purpose of an oncology dose-finding trial for novel anticancer agents has been shifting from determining the maximum tolerated dose (MTD) to identifying an optimal dose (OD) that is tolerable and therapeutically beneficial for subjects in subsequent clinical trials. In 2022, the FDA Oncology Center of Excellence initiated Project Optimus to reform the paradigm of dose optimization and dose selection in oncology drug development and issued a draft guidance. The guidance suggests that dose-finding trials include randomized dose-response cohorts of multiple doses and incorporate information on pharmacokinetics (PK) in addition to safety and efficacy data to select the optimal dose. Furthermore, PK information could be a quick alternative to efficacy data to predict the minimum efficacious dose and decide the dose assignment. This article proposes a model-based design for dose optimization with a randomization scheme based on PK outcomes in oncology. A simulation study shows that the proposed design has advantages compared to the other designs in the percentage of correct OD selection and the average number of patients assigned to OD in various realistic settings.

**China Dose Confirmation Designs in Early Phase Oncology Trials**

♦*Lusha Wang*

Bayesian optimal interval (BOIN) design, as a innovative model-assisted design, has received many attention from sponsors and clinical researchers at home and abroad because of its simple operation characteristics, easy understanding, and better performance. In this presentation, in combination with BOIN design, briefly summarize the operational characteristics, approach (using process) and relevant use scenarios and application cases of BOIN design from the statistical perspective in China dose confirmation designs in early phase oncology trials.

**Options and Challenges in Early Oncology Designs**

♦*Wentian Guo*1.

1Astrazeneca R&D China

The landscape of oncology drug development is dynamic, with novel treatments emerging rapidly. Efficient and robust early-phase designs are crucial for evaluating tolerability/toxicity and preliminary efficacy of investigational products. In this context, several considerations come into play, including but not limited to robust but fast efficient way to explore a range of doses, how to utilize efficacy, pharmacokinetics/pharmacodynamics and biomarkers for decision making, when and how to add backfills, how to consider long-term safety issues, and how to address questions raised by Project Optimus. In summary, early oncology designs must strike a balance between efficiency, patient safety, and robust decision-making.

**MC-Keyboard: A Practical Phase I Trial Design for Targeted Therapies and Immunotherapies Integrating Multiple-Grade Toxicities**

♦*Liyun Jiang*1, *Zhulin Yin*2, *Fangrong Yan*1 *and Ying Yuan*3.

1China Pharmaceutical University， 2Clinical Trials and Statistics Unit, Institute of Cancer Research， 3The University of Texas MD Anderson Cancer Center

Introduction: In targeted therapies and immunotherapies, the occurrence of low-grade (eg, grade 1-2) toxicities (LGT) is common, while dose-limiting toxicities (DLT) are relatively rare. As a result, conventional phase I trial designs, solely based on DLTs and disregarding milder toxicities, are problematic when evaluating these novel therapies.

Methods: To address this issue, we propose a novel phase I design called a multiple-constraint keyboard (MC-Keyboard) that integrates multiple toxicity constraints, accounting for both DLT and LGT, for precise dose escalation and de-escalation, and identification of the maximum tolerated dose (MTD). As a model-assisted design, an important feature of MC-Keyboard is that its dose-escalation or de-escalation rule can be pretabulated and incorporated into the trial protocol before the initiation of the trial, greatly simplifying its implementation.

Results: The simulation study showed that the MC-Keyboard had high accuracy in identifying the MTD and is safer than some existing designs.

Conclusion: The MC-Keyboard provides a novel, simple, and safe approach to assessing safety and identifying the MTD for targeted therapies and immunotherapies."

**Session 24CHI18: Analyses of Complex Survival Data**

**Parametric Model Checking Based on Wighted Residual Processes with Diverging Number of Predictors**

♦*Yue Hu*

The classical integrated conditional moment test will suffer from dimension problem in diverging dimension scenarios for model checking. To solve this problem, we propose a new test based on weighted residual processes for testing the parametric form of mean function in regression models. The asymptotic properties of the test statistic are investigated under the null and alternative hypotheses when the dimension is treated as a divergent number when the sample size tends to infinity. Our test can detect some classes of local alternative hypotheses converging to the null hypothesis at the parametric rate . As the limiting null distribution of our test is not asymptotically distribution-free, we use a smooth residual bootstrap to approximate the limiting null distribution. Simulations are conducted to examine the performance of our methodology. An analysis of real data is shown for illustration.

**Semiparametric Transformation Nonmixture Cure Models for Interval-Censored Failure Time Data with Mismeasurement Covariates**

♦*Rong Liu*1, *Yichen Lou*2, *Lei Ge*3,*Chunjie Wang*1

1Changchun University of Technology, 2Jinlin University, 3North East of Normal University.

This article discusses regression analysis of interval-censored failure time data in the presence of a cured subgroup or mismeasured covariates. Such data often occur in many areas, including epidemiological studies, medical studies, and social sciences. In this article, we discuss EM algorithm for interval-censored data with a cured subgroup or mismeasured covariates arising from the linear transformation model, a simulation-extrapolation approach is developed for the estimation. Furthermore, we show that the resulting estimators of regression parameters are consistent, asymptotically normal and semiparametrically efficient. An extensive simualtion study is conducted and suggests that the proposed approach works well in practical situations. Finally the method is applied to an example that motivated this study.

**Exploring Causal Effects of Hormone- and Radio-Treatments in An Observational Study of Breast Cancer under Semi-Competing Risks Setting**

♦*Tonghui Yu*1, *Mengjiao Peng*2, *Yifan Cui*3, *Elynn Chen*4, *Chixiang Chen*5

1Nanyang Technological University, Singapore, 2East China Normal University, China, 3Zhejiang University, China, 4New York University, USA, 5Tulane university

Breast cancer patients after surgery may suffer from relapse or death in the duration of follow-up. This phenomenon, known as semi-competing risk, demands advanced statistical tools for unbiased analysis. Despite progress in estimation and inference within semi-competing risks regression, its application to causal inference is still in its early stages. This article aims to establish a frequentist and semi-parametric framework based on copula models, facilitating valid causal inference, net quantity estimation, and sensitivity analysis for unmeasured factors under right-censored semi-competing risks data. We also study the non-parametric identification and propose novel procedures to enhance parameter estimation. We apply the proposed framework to a breast cancer study and detect the time-varying causal effects of hormone- and radio-treatments on patients' relapse-free survival and overall survival.

**A New and Unified Method for Regression Analysis of Interval-Censored Failure Time Data under Semiparametric Transformation Models with Missing Covariates**

♦*Yichen Lou*1, *Yuqing Ma*1, *Mingyue Du*1

1Jilin University

This paper discusses regression analysis of interval-censored failure time data arising from semiparametric transformation models in the presence of missing covariates. Although some methods have been developed for the problem, they either apply only to limited situations or may have some computational issues. Corresponding to these, we pro- pose a new and unified two-step inference procedure that can be easily implemented using the existing or standard software. The proposed method makes use of a set of working models to extract partial information from incomplete observations and yields a consistent estimator of regression parameters assuming missing at random. An extensive simulation study is conducted and indicates that it performs well in practical situations. Finally, we apply the proposed approach to an Alzheimer’s Disease study that motivated this study.

**Session 24CHI128: Some Topics on Probability and Statistics**

**FDR Control for Additive Hazards Model via Data Splitting**

♦*Jinzhao Yu*

School of Statistics and Mathematics, Zhongnan University of Economics and Law

The additive hazards model is one of the most popular semiparametric models in survival analysis and has been studied extensively for the modeling issues of survival data. However, the most existing results on variable selection for the framework of survival analysis have not considered the control of the false positives. In this paper, we proposed the covariate selection method via data splitting to control the false discovery rate for the additive hazards model. We construct the upper bound for our method under the finite sample condition. Furthermore under regularity conditions, we also discuss the asymptotic results for the proposed method regarding the FDR control theoretically. Simulation and real data analysis verify that our method can control the FDR while enjoy the high power.

**Variable Selection in Expectile Regression via Broken Adaptive Ridge**

♦*Mingyuan Shao*

School of Statistics and Mathematics, Zhongnan University of Economics and Law

Expectile regression is known to be efficient in detecting heteroscedasticity in high-dimensional data. Since it employs a smooth differentiable loss, the computational effort involved is considerably alleviated and theoretical analysis is also more amenable. In this paper, we consider a penalized expectile regression via broken adaptive ridge method. A generalized coordinate descent algorithm is proposed to solve the reweighted L2 penalty problem. The oracle property of the BAR estimator is established. Furthermore, we show that the BAR estimator also has a natural group structure. Numerical simulations and empirical analysis verify that our method performs well and has significant advantages, especially in heterogeneous regression settings.

**Early Warning of Information Casecades Based on Burst Branching Estimation**

*Yanqing Wang*1, ♦*Jinling Liu*1, *Dianni Wang*1.

1Zhongna University of Economics and Law.

Network information cascades refer to the phenomenon where specific information triggers a chain reaction of similar behaviors or viewpoints in others, leading to an explosive growth of information. Based on the local influence analysis method, we propose an Burst Branching Estimation within the branching process framework to identify potential information cascades from a multitude of public opinions. Through using Weibo online public opinion data as an empirical study, we have verified the effectiveness of this method.

**Limit Theorems for a Supercritical Two-type Decomposable Branching Process in a Random Environment**

*Yanqing Wang*1, ♦*Dianni Wang*1, *Jinling Liu*1, *Quansheng Liu*1

1Zhongnan University of Economics and Law

Let $Z\_n=(Z\_n^{(1)},Z\_n^{(2)})$ be a two-type decomposable branching process in an independent and identically distributed random environment, where a type $1$ particle may produce particles of types 1 and 2, while a type 2 particle can only give birth to type 2 particles. We consider asymptotic properties of this process in the supercritical case. Because $Z\_n^{(1)}$ is an usual single-type branching process in a random environment, we only consider $Z\_n^{(2)}$. First, under some moment conditions, we find a suitable factor $\Pi\_n$ such that the normalized population size $W\_n=\frac{Z\_n^{(2)}}{\Pi\_n}$ converges almost surely to a finite random variable $W$, and provide a decomposition expression and a non-degeneracy condition of $W$. Second, we give conditions under which $(W\_n)$ is convergent in $L^p$ for $p\geq1$, and bounded in $L^p$ for $0<p<1$. Finally, we establish a central limit theorem for $\log Z\_n^{(2)}$.

**Session 24CHI1: Advanced Methods for Data with Heterogeneity**

**Change Point Detection for High-Dimensional Linear Models: A General Tail-Adaptive Approach**

♦*Bin Liu*1, *Zhengling Qi*2, *Xinsheng Zhang*1 *and Yufeng Liui*3.

1Fudan University，2George Washington University， 3University of North Carolina at Chapel Hill

We study the change point detection problem for high-dimensional linear regression models. In this work, we propose a novel tail-adaptive approach for simultaneous change point testing and estimation. The method is built on a new loss function which is a weighted combination between the composite quantile and least squared losses, allowing us to borrow information of the possible change points from both the conditional mean and quantiles. For the change point testing, based on the adjusted L2-norm aggregation of a weighted score CUSUM process, we pro-pose a family of individual testing statistics with different weights to account for the unknown tail structures. Through a combination of the individual tests, a tail-adaptive test is further constructed that is powerful for sparse alternatives of regression coefficients’ changes under various tail structures. For the change point estimation, a family of argmax-based individual estimators is proposed once a change point is detected. In theory, for both individual and tail-adaptive tests, bootstrap procedures are proposed to approximate their limiting null distributions. Under some mild conditions, we justify the validity of the new tests in terms of size and power under the high-dimensional setup. The corresponding change point estimators are shown to be rate optimal up to a logarithm factor.

**Oracle Change Point Detection and Localizaiton For High-Dimensional Graphical Models**

♦*Xianru Wang*1 , *Bin Liu*2, *Xinsheng Zhang*2*，Yufeng Liui*3.

1The Chinese University of Hong Kong， 2Fudan University，3University of North Carolina at Chapel Hill

In this paper, we consider efficient multiple change point detection and localizaiton for high dimensional graphical models. Based on the Nodewise Lasso estimation of gragh structure, a two-step change point estimation methods and two algorithms are proposed. The estimation method is constructed based on loss function and we develop the Refined Dynamic Programming Algorithm (RDPA) and the Refined Wild Binary Segmentation (RWBS) to detect change points. Under the mild model assumptions, we present the theoretical results of both two algorithms, including the estimation consistency of change point number, the oracle estimation of change point locations, and the estimation of the underlying coefficient vectors with optimal covergence rate. At last, extensive numerical results and the application to an Alzheimer's disease dataset demonstrate the computational efficiency and the statistical accuracy of our proposed methods.

**Online Multiple Changepoints Detection with False Discovery Rate Control**

*Xiaolong Cui*1, *Haoyu Geng*1,♦*Haojie Ren*2, *Zhaojun Wang*1*，Changliang Zou*1.

1Nankai University，2Shanghai Jiao Tong University

Technological advances have led to the emergence of an increasing number of applications requiring analysis of datastreams, that are characterized by an indefinitely long and time-evolving sequence. In such applications, the status of a stream can alternate, possibly many times, between a regular status and an irregular status. Consequently, it is necessary to develop statistical methodologies that constantly detect multiple changepoints in an online manner. While we may employ conventional methods of sequential change detection to trigger signals after the change occurs, no online procedure is available to quantify the uncertainty of the detected changes. In this work, we fill this gap by framing online multiple changepoint detection into an online multiple testing problem and proposing a new framework to test the null hypothesis that there is no change between neighboring signalled points. To obtain valid p-values for online multiple testing, we propose a data-fission-based procedure that is a simple yet effective way of dealing with post-detection uncertainty quantification. It is shown that popular online false discovery rate (FDR) control methods with those p-values can achieve finite-sample FDR control. The advantage of the proposed method is demonstrated via simulation studies and a data example.

**Tensor Additive Quantile Regression**

♦*Wenqi Lu*1, *Zhongyi Zhu*2,*Heng Lian*3

1School of Statistics and Data Science, 2Nankai University, School of Management Fudan University, 3Department of Mathematics City University of Hong Kong

Tensor analysis becomes popular because of its extensive use in fields with high-order structured data. However, most tensor regression studies focus on linear regression model with Gaussian errors, which is restricted in real applications. This paper considers a nonparametric additive quantile regression model with tensor covariates, which accommodates non-linearity and non-Gaussian errors. To estimate the nonparametric additive component functions, we approximate them by spline series expansion and arrange the spline coefficients into a tensor, which induces a linear tensor regression with transformed predictors. Moreover, in order to tackle the high-dimensionality which often arises in modern tensor regression applications, we introduce a sparse assumption which assumes the number of nonzero additive components is smaller than the sample size, and apply group lasso to encourage the sparsity. For both fixed-dimensional and high-dimensional sparse models, we establish asymptotic properties for the proposed estimators and provide alternating update algorithms. The efficiency of the algorithms is shown in simulations, and applications to a stock market data set and a head pose estimation problem.

**Session 24CHI24: Complex Medical Data Analysis**

**Isease-Specific Enhancement for Biological Networks Improves Drug Repurposing**

♦*Peifeng Ruan.*

Biological network-based approaches have been proven to be helpful in repurposing drugs, resulting in significant cost and time reduction in drug development. However, the gene networks used are not disease-specific and cannot represent true gene interactions for a specific disease. Utilizing an established human gene network, we propose a general framework that enhances the network for a specific disease by using disease-specific omics data through a diffusion process. Then the enhanced disease-specific gene network is integrated with networks of interconnected drugs and diseases and helps repurposing drugs. Applications of omics data from The Cancer Genome Atlas (TCGA) project suggest that the enhanced disease-specific networks achieve a significant improvement in drug repurposing tasks than the general human gene network.

**Semi-Parametric Panel Count Model, with Applications to Signal Detection in Post-Market Drug Surveillance Systems**

♦*Yizhao Zhou*1, *Ao Yuan*2*，Ming Tan*2.

1AstraZeneca Global R&D (China)， 2Georgetown University

Panel count data occurs in various applications ranging from biomedical research to business, such as the number of accidents, product defects, and insurance claims. Under the US FDA system, millions of reported adverse events (AEs) associated with thousands of drugs are monitored in the post-market drug safety surveillance systems worldwide. Evaluating the AEs of the associated drugs is of important public health concern and motivates our method. One statistical challenge in such systems is handling excessive zero AE counts. Most existing methods utilize Poisson counts models that can not incorporate covariates nor adequately account for the excessive zero counts. This article proposes a novel semi-parametric, non-homogeneous panel count model to detect AE signals by accounting for covariates, background AE occurrences, and excessive zero counts. The model is estimated using the Expectation-Maximization (EM) algorithm iteratively, wherein, for each M-step, the maximization of the non-parametric component is reformulated as an optimization problem as in the isotonic regression. The strong consistency and the asymptotic distributions of the estimators are formally derived. We conduct extensive simulation studies to evaluate the finite sample performance of the proposed method and to demonstrate the apparent advantage of the proposed method in signal detection with high power, high specificity, and sensitivity. We apply the method to a WHO VigiBase dataset to detect the AE's signals, which is an application of the proposed method.

**Optimal Weighted Random Forests**

*Xinyu Chen*1, ♦*Dalei Yu*2, *Xinyu Zhang*3.

1University of Science and Technology of China, 2Xi'an Jiaotong University，3Academy of Mathematics and Systems Science, Chinese Academy of Sciences

The random forest (RF) algorithm has become a very popular prediction method for its great flexibility and promising accuracy. In RF, it is conventional to put equal weights on all the base learners (trees) to aggregate their predictions. However, the predictive performances of different trees within the forest can be very different due to the randomization of the embedded bootstrap sampling and feature selection. In this paper, we focus on RF for regression and propose two optimal weighting algorithms, namely the 1 Step Optimal Weighted RF (1step-WRF$\_\mathrm{opt}$) and 2 Steps Optimal Weighted RF (2steps-WRF$\_\mathrm{opt}$), that combine the base learners through the weights determined by weight choice criteria. Under some regularity conditions, we show that these algorithms are asymptotically optimal in the sense that the resulting squared loss and risk are asymptotically identical to those of the infeasible but best possible weighted RF. Numerical studies conducted on real-world data sets and semi-synthetic data sets indicate that these algorithms outperform the equal-weight forest and two other weighted RFs proposed in existing literature in most cases.

**DiffGR: Detecting Differentially Interacting Genomic Regions from Hi-C Contact Maps**

♦*Huiling Liu*1, *Wenxiu Ma*2

1South China University of Technology, 2University of California, Riversid

Recent advances in Hi-C techniques have allowed us to map genome-wide chromatin interactions and uncover higher-order chromatin structures, thereby shedding light on the principles of genome architecture and functions. However, statistical methods for detecting changes in large-scale chromatin organization such as topologically-associating domains (TADs) are still lacking. We proposed a new statistical method, DiffGR, for detecting differentially interacting genomic regions at the TAD level between Hi-C contact maps. We utilized the stratum-adjusted correlation coefficient to measure similarity of local TAD regions. We then developed a nonparametric approach to identify statisticallysignificant changes of genomic interacting regions. Through simulation studies, we demonstrated that DiffGR can robustly and effectively discover differential genomic regions under various conditions. Furthermore, we successfully revealed cell type-specific changes in genomic interacting regions in both human and mouse Hi-C datasets, and illustrated that DiffGR yielded consistent and advantageous results compared with state-of-the-art differential TAD detection methods.

**Session 24CHI45: Moderate Deviation and Inference for Independent and Dependent Data**

**Some Results on Probabilities of Moderate Deviation**

*Deli Li*1, *Yu Miao*2，♦*Yongcheng Qi*3.

1Lakehead University， 2Henan Normal University， 3University of Minnesota Duluth

Consider a sequence of independent and identically distributed non-degenerate real-valued random variables with a finite second moment. We obtain precise asymptotic estimates for the probabilities of moderate deviations of the partial sums. Those known results under assumption of high moments in the literature depend on the variance of the underlying distribution only. The moderate deviation results established in this paper depend on both the variance and the asymptotic behavior of the tails of the underlying distribution.

**Moderate Deviation Principle of Modularity in Network**

*Qin Yin*1, *Yu Miao*1, *Zhen Wang*1*，* ♦*Guangyu Yang*2.

1Henan Normal University，2Zhengzhou University

Complex networks describe a wide range of systems in nature and society. Frequently cited examples include the cell, a network of chemicals linked by chemical reactions, and the Internet, a network of routers and computers connected by physical links. Traditionally these systems have been modeled as random graphs and one of the most relevant features is community structure. In the present paper, we study a specific partition of a given network and establish the moderate deviation principle of modularity for the partition when the size of the network gets large.

**Cramer's Moderate Deviations for Martingales with Applications**

♦*Xiequan Fan*1, *Qiman Shao*2

1Northeastern University at Qinhuangdao, 2Southern University of Science and Technology

Let $(\xii,\mathcal{F}\_i)\_{i\geq1}$ be a sequence of martingale differences. Set $Xn=\sum\_{i=1}^n \xii $and$\langleX\ranglen=\sum\_{i=1}^n\mathbf{E}(\xi\_i^2|\mathcal{F}\_{i-1}).$ We present some Cram\'er's moderate deviation expansions for $\displaystyle\ mathbf{P}(X\_n/\sqrt{\langleX\rangle\_n}\geqx)$and$\displaystyle\mathbf{P}(X\_n/\sqrt{\mathbf{E}X\_n^2}\geq x)$ as $n\to\infty.$ Our results extend the classical Cram\'{e}r result to the cases of normalized martingales $Xn/\sqrt{\langle X\rangle\_n} $and standardized martingales $Xn/\sqrt{\mathbf{E} X\_n^2}$, with martingale differences satisfying the conditional Bernstein condition. Applications to elephant random walks and autoregressive processes are also discussed.

**Inference Via Robust Optimal Transportation: Theory and Methods**

*Ma Yiming*1, ♦*Hang Liu*1, *Davide La Vecchia*2*， Matthieu Lerasle*3

1University of Science and Technology of China, 2University of Geneva， 3ENSAE Paris

Optimal transportation theory and the related $p$-Wasserstein distance ($Wp$, $p\geq 1$) are widely-applied in statistics and machine learning. In spite of their popularity, inference based on these tools has some issues. For instance, it is sensitive to outliers and it may not be even defined when the underlying model has infinite moments. To cope with these problems, first we consider a robust version of the primal transportation problem and show that it defines the robust Wasserstein distance, $W^{(\lambda)}$, depending on a tuning parameter $\lambda > 0$. Second, we illustrate the link between $W\_1$ and $W^{(\lambda)}$ and study its key measure theoretic aspects. Third, we derive some concentration inequalities for $W^{(\lambda)}$. Fourth, we use $W^{(\lambda)}$ to define minimum distance estimators, we provide their statistical guarantees and we illustrate how to apply the derived concentration inequalities for a data driven selection of $\lambda$. Fifth, we provide the dual form of the robust optimal transportation problem and we apply it to machine learning problems (generative adversarial networks and domain adaptation). Numerical exercises provide evidence of the benefits yielded by our novel methods.

**Session 24CHI108: Statistical Analysis of Complex Data**

**Testing latent classes in gut microbiome data using beta-binomial regression models**

*Qinglin Wang*1, ♦*Peng Ye*1，*Hua He*2.

1University of International Busines and Economics， 2Tulane University

In this talk, we develop a new test for detecting zero-inflation resulting from a latent class of subjects with structural zeros in a beta-binomial (BB) regression model by directly comparing the amount of observed zeros with what would be expected under the BB model. A closed form of the test statistic as well as its asymptotic properties are derived based on estimating equations. Simulation studies are conducted to investigate the performance of the new test and compare it with the classical Wald, likelihood ratio, and score tests. The tests are also applied to human gut microbiome data to test latent class in microbial genera.

**High-Dimensional Partially Linear Functional Cox Models**

♦*Xin Chen*, *Hua Liu*, *Jiaqi Men，Jinhong You*.

As a commonly employed method for analyzing time-to-event data involving functional predictors, the functional Cox model assumes a linear relationship between the functional principle component (FPC) scores of the functional preditors and hazard rates. However, in practical scenarios such as our study on the survival time of kidney transplant recipients, the assumption does not hold. To address this limitation, this paper introduces a high-dimensional partially linear functional Cox model (PLFC), which not only allows for non-linear effects of functional predictors on the response, but also accommodates the diverging numbers of scalar predictors and FPCs as the sample size increases. We employ the group SCAD method for selecting relevant scalar predictors and FPCs and incorporate B-splines for a smoothed estimate of the non-linear effect. The finite sample performance of the estimates is evaluated through simulation studies. Applying the model to a kidney transplant dataset from the Organ Procurement and Transplantation Network also enables inference on the nonlinear functional relationships between FPC scores of eGFR and hazard rates, as well as selecting important predictors for long-term survival time.

**Real-Time Inference for Streaming Survival Data from Multiple Heterogeneous Studies with A Cure Fraction**

♦*Bo Han*1, *Niansheng Tang*1, *Liuquan Sun*2 *， Van Keilegom Ingrid*3.

1Yunnan University， 2Academy of Mathematics and Systems Science, Chinese Academy of Sciences， 3KU Leuven

In survival analysis, it often happens that a certain fraction of subjects will never experience the event of interest and thus they are considered to be cured. This paper proposes a novel procedure to draw real-time inference for covariate effects on survival with a cure fraction. For the promotion time cure model, we first investigate an online method by combining the likelihood function from current data with the confidence density function of summary statistics from historical data. It enables borrowing of strength from summary-level information, thereby relaxing the assumption of model homogeneity among data batches and enjoying computational efficiency. The consistency and weak convergence of the online estimator are established and it is shown to achieve statistical efficiency. We then propose an online data fusion method for streaming data from multiple heterogeneous studies, which synthesizes inferential information from all studies to make more effective inference than from any study alone. The inference procedure is easy to implement using standard statistical software and computationally fast without involving resampling. Our methods are illustrated via simulation studies and an application to breast cancer data.

**Efficient Estimation and Optimal Treatment Allocation in Adaptive Clinical Trials**

♦*Wei Zhang*1, *Zhiwei Zhang*2，*Aiyi Liu*3.

1Academy of Mathematics and Systems Science, CAS， 2Biostatistics Innovation Group, Gilead Sciences， 3National Institute of Child Health and Human Development, National Institutes of Health

When the sample size of preliminary trial data is small, an adaptive design is commonly used. We consider the problem of efficient estimation and optimal treatment allocation in adaptive (two or multi-stage) clinical trials. Under both the conventional covariate-independent randomization (CIR) and new covariate-dependent randomization (CDR), we develop efficient estimators for treatment effect by incorporating covariate information. Accordingly, we derive the optimal allocation ratio by maximizing the design efficiency of an adaptive randomized trial, assuming that an efficient estimator will be used for analysis. We show that treatment effect estimators are consistent, asymptotically normal, and (nearly) efficient under both CIR and CDR, and derive the optimal propensity score by maximizing the design efficiency of a CDR trial. Real world examples and simulation results demonstrate that the proposed designs can produce substantial efficiency improvements in realistic settings.

**Session 24CHI2: Advanced Methods in Large-scale Statistical Analysis**

**Profiled Transfer Learning for High Dimensional Linear Model**

♦*Ziqian Lin*, *Junlong Zhao,* *Fang Wang**， Hansheng Wang.*

We develop here a novel transfer learning methodology called Profiled Transfer Learning (PTL). The method is based on the \textit{approximate-linear} assumption between the source and target parameters. Compared with the commonly assumed \textit{vanishing-difference} assumption and \textit{low-rank} assumption in the literature, the \textit{approximate-linear} assumption is more flexible and less stringent. Specifically, the PTL estimator is constructed by two major steps. Firstly, we regress the response on the transferred feature, leading to the profiled responses. Subsequently, we learn the regression relationship between profiled responses and the covariates on the target data. The final estimator is then assembled based on the \textit{approximate-linear} relationship. To theoretically support the PTL estimator, we derive the non-asymptotic upper bound and minimax lower bound. We find that the PTL estimator is minimax optimal under appropriate regularity conditions. Extensive simulation studies are presented to demonstrate the finite sample performance of the new method. A real data example about sentence prediction is also presented with very encouraging results.

**Quasi-Newton Updating for Large-Scale Distributed Learning**

♦*Shuyuan Wu*1, *Danyang Huang*2*， Hansheng Wang*3.

1Shanghai University of Finance and Economics， 2Renmin University of China， 3Peking University

Distributed computing is critically important for modern statistical analysis. Herein, we develop a distributed quasi-Newton (DQN) framework with excellent statistical, computation, and communication efficiency. In the DQN method, no Hessian matrix inversion or communication is needed. Thisconsiderably reduces the computation and communication complexity of the proposed method. Notably, relatedexisting methods onlyanalyze numerical convergence and require a diverging number of iterations to converge. However, we investigate the statistical properties of the DQN method and theoretically demonstrate that the resulting estimator is statistically efficient over a small number of iterations under mild conditions. Extensive numerical analyses demonstrate the finite sample performance.

**Communication-Efficient Distributed Sparse Learning with Oracle Property and Geometric Convergence**

♦*Jiyuan Tu*1, *Weidong Liu*2*， Xiaojun Mao*2.

1Shanghai University of Finance and Economics， 2Shanghai Jiao Tong University

This article introduces two highly efficient distributed non-convex sparse learning algorithms. Our approach accommodates non-convexity in both the loss function and penalty, acknowledging the potential non-uniqueness of local minimizers due to the inherent non-convexity. The development of an algorithm that ensures convergence to a locally minimal solution with desired statistical properties becomes imperative in this context. To overcome this challenge, we propose a strategy involving the relaxation of non-convexity in the penalty function through a local linear approximation. Addressing non-convexity in the loss function, we employ the proximal homotopy method, initiating the process with a relatively large regularization parameter that is gradually shrunk towards the target parameter. Theoretical considerations form the cornerstone of our work, in which we provide an explicit statistical rate for the approximated solution in each iteration. Our contributions extend to establishing oracle properties and demonstrating asymptotic normality, marking a significant advancement as the first asymptotic normality result for the approximated solution yielded by the algorithm. Consequently, our theoretical framework offers valuable insight into setting the optimization error, especially when undertaking statistical inference for the approximate estimator. In terms of computational efficacy, our algorithm exhibits a geometric convergence rate within each middle loop. We substantiate the validity of our theory through a comprehensive presentation of numerical results. This multifaceted analysis underscores the applicability of our proposed algorithms, consolidating their credibility in both the theoretical and computational domains.

**Mini-Batch Gradient Descent with Buffer**

♦*Haobo Qi*1, *Du Huang*2, *Yingqiu Zhu*3, *Danyang Huang*4 *and Hansheng Wang*5.

1Beijing Normal University， 2Matpool, Timestamp Information Technology LLC， 3University of International Business and Economic， 4Renmin University of China 5Peking University

In this paper, we studied a buffered mini-batch gradient descent (BMGD) algorithm for training complex model on massive datasets. The algorithm studied here is designed for fast training on a GPU-CPU system, which contains two steps: the buffering step and the computation step. In the buffering step, a large batch of data (i.e., a buffer) are loaded from the hard drive to the graphical memory of GPU. In the computation step, a standard mini-batch gradient descent (MGD) algorithm is applied to the buffered data. Compared to traditional MGD algorithm, the proposed BMGD algorithm can be more efficient for two reasons. First, the BMGD algorithm uses the buffered data for multiple rounds of gradient update, which reduces the expensive communication cost from the hard drive to GPU memory. Second, the buffering step can be executed in parallel so that the GPU does not have to stay idle when loading new data. We first investigate the theoretical properties of BMGD algorithms under a linear regression setting. The analysis is then extended to the Polyak-Lojasiewicz loss function class. The theoretical claims about the BMGD algorithm are numerically verified by simulation studies. The practical usefulness of the proposed method is demonstrated by three image-related real data analysis.

**Session 24CHI12: Advances in Statistical Learning Inference with High Dimensional Data**

**Detecting Epidemic Changes Through Shifted Maximum Subarray Analysis**

*♦Yue Niu1, Ning Hao1, Tim Mo1*， *Han Xiao2.*

1University of Arizona，2Rutgers University

Circular binary segmentation (CBS) is a popular approach to detecting epidemic changes with applications in array comparative genomic hybridization (aCGH) or next-generation sequencing data for copy number variation detection. At each step of CBS, an epidemic test is applied to determine whether the current segment is split to two segments. CBS improves the regular binary segmentation method in the power of detecting short segments, with an additional computation cost. In this talk, we introduce a fast method in computing the likelihood ratio test statistic for the epidemic test, which improves the computation efficiency of the CBS method. The new method is based on a well-known problem called maximum subarray problem, which can be solved efficiently via dynamic programming. We investigate a variant of maximum subarray problem, called shifted maximum subarray (SMS) problem, which studies the maximum subarray when all the values in the original array are shifted by a tuning parameter. We characterize and implement the calculation of the full solution path of the SMS problem when the tuning parameter varies. Moreover, we show that the epidemic test statistic can be recovered from the solution of SMS problem. As a result, we give a faster CBS algorithm.

**Robust Mendelian Randomization Method Accounting for Idiosyncratic and Correlated Pleiotropy with Applications to Stroke Outcomes.**

♦*Qing Cheng1, Wenxin Xu1, Chan Wang2* and *Yanyan Zhao3.*

1Southwestern University of Finance and Economics， 2New York University Grossman School of Medicine， 3Shandong University

Mendelian randomization (MR) serves as a valuable tool for investigating causal relationships between exposures and disease outcomes in observational studies. However, MR methods, operating under classical assumptions, may yield biased estimates and

inflated false-positive causal relationships when faced with realistic and complex correlated horizontal pleiotropy (CHP). While numerous MR methods have emerged to address CHP effects, limited methods can effectively handle relatively large direct effects, commonly known as idiosyncratic pleiotropy. In response to this gap, we propose an efficient and Robust Mendelian Randomization method to account for Idiosyncratic and Correlated Pleiotropy, named RMR-ICP. Furthermore, our method employs paralleled Gibbs sampling to incorporate linkage disequilibrium structure, thereby enhancing statistical power. We demonstrate the robustness and efficiency of our method through extensive simulation studies and applications. Particularly, we apply RMR-ICP to study the effects of plasma proteins on stroke. Several notable associations are identified. For example, SELE has a positive causal effect on any stroke. An elevated BNP is associated with an increased risk of cardioembolic stroke, but not with other stroke subtypes. This offers a fresh perspective in the identification of plasma proteins associated with stroke.

**Global and Large-Scale Multiple Testing for High-Dimensional Quantile Effects**

♦*Yeqing Zhou*1, *Kai Xu*2, *Chen Xu*3 *and Liping Zhu*4.

1Tongji University, 2Anhui Normal University, 3University of Ottawa and 4Renmin University of China

In this paper, global and large-scale multiple testing procedures are developed to detect the significant covariates that contribute to the conditional quantiles of a response in the high-dimensional setting. We first introduce a new metric, the so-called nonparametric quantile correlation, to measure the departure of conditional quantile independence. The nonparametric quantile correlation is resistent to the heavy-tailed response and covariates, and is useful in analysing the nonlinear quantile relationships for the heterogeneous data. A test statistic for the global null hypothesis is developed based on the summation of squared marginal nonparametric quantile covariances, without assuming any specific functional form of the quantile regression. We establish its asymptotic normality as the dimension and sample size both go to infinity. To test the sig- nificance of each covariate simultaneously, we propose a data-driven multiple testing procedure via a sample-splitting strategy, which effectively controls both false discovery rate and false discovery proportion asymptotically. We illustrate the advantages of the proposed methods via extensive simulation studies and a gene expression dataset.

**Enhancing Copy Number Variation Detection with Deep Learning Techniques**

♦*Feifei Xiao1, Dayuan Wang1* and *Fei Qin*

1University of Florida，2National Institute of Health

Copy Number Variants (CNVs) are duplications or deletions of genome segments and are associated with risks of various human diseases, including cancer and neurodevelopmental disorders. Accurate CNV detection is crucial for understanding their roles in disease etiology. However, existing CNV calling methods often suffer from high false-positive rate and low accuracy. In this study, we present an innovative approach to refine CNV detection by leveraging deep learningtechniques. To construct high quality ground truth variants, we utilized a publicly available whole exome sequencing dataset from the publicly available Autism Genome Project. The dataset comprised samplesfrom over 1,100 trio families, enabling us to construct a large high-confidence variant set. Initial CNV calling was performed with series of state-of-the-art CNV detection tools (e.g., EXCAVATOR2, CORRseq). We then labeled the detected CNVs as true/false discoveries according to Mendelian Inheritance Rules with the family trios. The pre-specified CNV regions were selected and then subjected to develop a transfer learning model for CNV detection. We took strengths from contemporary deep learning models such as Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN) to capture complex patterns and dependency within the genomic data, thereby enhancing the accuracy of CNV detection. Comparing with existing CNV callers, our results indicate promising improvements in CNV detection accuracy. As such, this research contributes to the development of more reliable tools for structural variants study, which will advance our understanding of the role of CNVs in various diseases and ultimately improve clinical diagnostics.

**Session 24CHI56: Novel Statistical Methods with Applications in Medical and Financial Studies**

**Screen Then Select: A Strategy for Correlated Predictors in High-Dimensional Quantile Regression**

♦*Xuejun Jiang1*

1Southern University of Science and Technology

Strong correlation among predictors and heavy-tailed noises pose a great challenge in the analysis of ultra-high dimensional data. Such challenge leads to an increase in the computation time for discovering active variables and a decrease in selection accuracy. To address this issue, we propose an innovative two-stage screen-then-select approach and its derivative procedure based on a robust quantile regression with sparsity assumption. This approach initially screens important features by ranking quantile ridge estimation and subsequently employs a likelihood-based post-screening selection strategy to refine variable selection. Additionally, we conduct an internal competition mechanism along the greedy search path to enhance the robustness of algorithm against the design dependence. Our methods are simple to implement and possess numerous desirable properties from theoretical and computational standpoints. Theoretically, we establish the strong consistency of feature selection for the proposed methods under some regularity conditions. In empirical studies, we assess the finite sample performance of our methods by comparing them with utility screening approaches and existing penalized quantile regression methods. Furthermore, we apply our methods to identify genes associated with anticancer drug sensitivities for practical guidance.

**GPJM: Gaussian Process Joint Model for Integrative Analysis of Parkinson's Disease**

*Junxuan Chen1* and♦*Kai Kang1*

1Sun Yat-sen University

Patients affected by Parkinson's disease (PD) usually experience a range of symptoms, including movement disorder, sleep disturbance, cognitive impairment and brain structural changes, throughout clinical course. Current research focusing on a single domain (e.g., movement disorder) fails to display the full picture of the disease progression. In this paper, we integrate longitudinal mixed types of measurements from multiple domains (e.g., binary clinical symptoms, ordinal cognitive measures, continuous neuroimaging biomarkers) and uncover their relationship with the initiation of PD treatment using a Gaussian process joint model. The dependence structure between observed multi-domain biomarkers is characterized by several underlying unobserved Gaussian processes through a generalized factor model. Instead of assuming a certain parametric form a priori, the Gaussian processes, with its philosophy to let data speak for themselves, are expected to capture the possibly nonlinear and complex progression profiles of PD-related measurements. The obtained Gaussian processes are also incorporated into a varying coefficient Cox model so as to jointly monitor the longitudinal PD-related measurements and their time-varying effects on time-to- initiation of PD treatment. The application of our proposed method to the Parkinson’s Progression Markers Initiative (PPMI) dataset yields a comprehensive understanding of PD progression by synthesizing information from clinical, cognitive, and neuroimaging perspectives.

**Threshold Estimation in Proportional Mean Residual Life Model**

♦*Bing Wang1* and *Xinyuan Song2*

1Anhui University, 2The Chinese University of Hong Kong

The mean residual life model is vital for its ability to investigate the association between covariates and patient life expectancy. In certain circumstances, a patient’s lifespan may change when a covariate exceeds a particular threshold value, which is critical to predicting the patient’s life expectancy and preventing diseases. This study considers a threshold regression analysis of a proportional mean residual life model with a continuous thresholding variable. We construct martingale-based smoothed estimating equations to obtain parameter estimators, and establish the large-sample properties of the proposed estimators. Furthermore, we propose a supremum test to examine the existence of the threshold. Finally, we assess the finite-sample performance of the proposed method using simulation studies, and then apply the methodology to data from colorectal and breast cancer studies.

**Regression Analysis of Logistic Model with Latent Variables**

*Yuan Ye*1, *Zhongchun Liu* 2, ♦*Deng Pan*1 and *Yuanshan Wu*3.

1Huazhong University of Science and Technology， 2Renmin Hospital of Wuhan University，3Zhongnan University of Economics and Law

We propose a joint modeling approach to investigating the effects of social-psychological factors on the onset of depression. The proposed model comprises two components. The first one is a confirmatory factor analysis model that summarizes latent factors through multiple correlated observed variables. The second one is a logistic regression model that investigates the effects of observed and latent influence factors on the occurrence of depression. We develop a hybrid procedure based on the borrow-strength estimation procedure and the weighted score function to estimate the model parameters. The asymptotic properties of the proposed estimators are established. Simulation studies demonstrate that the method we proposed performs well. An application to a study concerning the social-psychological factors of depression is provided.

**Session 24CH74: Recent Advances in Statistical Methods for Analyzing Two-Phase Data and Genetic Data**

**Moving Beyond Population Variable Importance: Concept, Theory and Applications of Individual Variable Importance**

♦*Guorong Dai1, Lingxuan Shao1* and *Jinbo Chen2*

1Fudan University， 2University of Pennsylvania

In a nonparametric regression setting, we propose a novel concept of ''individual variable importance'', which refers to the relevance of some covariates with respect to an outcome variable among individuals with certain features. This concept holds practical importance for both risk assessment and association identification. For example, it can represent (i) the usefulness of expensive biomarkers in disease prediction for individuals at certain baseline risk, or (ii) age-specific associations between physiological indicators. We quantify the individual variable importance by a ratio parameter between two conditional mean squared errors. To infer this parameter we develop fully nonparametric estimators and establish their asymptotic properties. Our inferential approaches perform well in simulation studies. We further demonstrate our approaches through application to a real data set, showing a scientifically interesting result: the association between bodyshape and systolic blood pressure decays with increasing age. While our finding aligns with the existing medical literature based on standard parametric regression techniques, our analysis methods are more reliable because their validity is not affected by model misspecification. More importantly, the fully nonparametric nature equips the individual variable importance framework with broader applicability in contexts that go beyond traditional parametric modeling. It can capture complex relationships between variables that are likely to be missed by standard parametric interaction analyses.

**Mutually Exclusive Spectral Biclustering and Its Applications in Cancer Subtyping**

♦*Min Yuan*.

Anhui Medical University.

Many soft biclustering algorithms have been developed and applied to various biological and biomedical data analyses. However, few mutually exclusive (hard) biclustering algorithms have been proposed to identify disease or molecular subtypes based on genomic or transcriptomic data. In this study, we developed a novel mutually exclusive spectral biclustering (MESBC) algorithm to detect mutually exclusive biclusters. MESBC simultaneously detects relevant features (genes) and corresponding patient subgroups and, therefore, automatically uses the signature features for each subtype to perform the clustering. Our simulations revealed that MESBC provided superior accuracy in detecting pre-specified biclusters compared with the non-negative matrix factorization (NMF) and Dhillon’s algorithm, particularly in very noisy data. Further analysis of the algorithm on real datasets obtained from the TCGA database showed that MESBC provided similar or more accurate overall survival prediction in patients with breast and lung cancer when compared to the existing, gold-standard subtypes for breast (PAM50) and lung cancer (integrative clustering). In the TCGA lung cancer patients, MESBC detected two clinically relevant, rare subtypes that were not detected by other biclustering or integrative clustering algorithms. Therefore, MESBC could potentially be used as a risk stratification tool to optimize the treatment for the patient, improve the selection of patients for clinical trials, and contribute to the development of novel therapeutic agents.

**Adaptive Stratified Sampling Design in Two-Phase Studies for Average Treatment Effect Estimation**

♦*Min Zeng1, Jinfeng Xu2* and *Hong Zhang1.*

1University of Science and Technology of China， 2City University of HongKong

Observational causal inference studies suffer from numerous confounding effects by greatly distorting average treatment effect (ATE) estimation if ignored. Information of some confounders, such as genetic biomarkers and medical imaging, is prohibitively expensive to obtain. Two-phase study is a resource-efficient solution to this problem. In such study, outcome, treatment, and inexpensive confounders are measured for a large number of subjects in the first phase; costly confounder measurements are then collected for a limited number of subjects in the second phase. An efficient statistical design is essential in controlling the cost arising from the second phase. In this paper, we propose an adaptive stratified sampling design ASCC, which minimizes the variance of the ATE estimator with a given second-phase sample size. ASCC begins with gathering costly confounder measures for randomly selected pilot data, which are used to develop a stratification strategy and determine the sampling probabilities of strata. The resulting stratification and sampling strategy is applied to all first-phase subjects to determine the second-phase subjects with costly confounders measures. We rigorously show that ASCC produces a more efficient ATE estimator compared with the existing sampling designs with strata being pre-fixed. Finite sample properties of ASCC are evaluated through simulation studies, which are shown to be more efficient than pre-fixed strata sampling designs, with efficiency gains ranging from 20\% to 30\% in our simulation situations. The desired finite sample properties of ASCC are further confirmed through the application to real examples.

**Methods for Identifying Mutually Exclusive Genes and Their Application in Breast Cancer Data**

♦*Zeyu Zhang*

Shijiazhuang Tiedao University

Background: The process of somatic variation in cancer can be viewed as a Darwinian evolution process. Studies have found that mutations in genes regulating the same pathogenic pathway often exhibit mutually exclusive relationships in cancer cells. Breast tumors are a heterogeneous group, with different molecular subtypes of breast cancer patients showing variations in treatment responses and prognoses. This study analyzed breast cancer data from The Cancer Genome Atlas (TCGA) to identify driver genes.

Methods: This study introduces two methods for identifying mutually exclusive genes. The first method, stratified mutually exclusive gene finding (SME), utilizes the Cochran-Mantel-Haenszel test (CMH test) to determine conditional exclusivity of genes. It combines distance ranking and linear search to identify candidate gene sets with potential mutual exclusivity. To address the problem of multiple mutual exclusivity tests, the Benjamini-Hochberg procedure is employed to control the false discovery rate (FDR). The second method is a forward stagewise selection algorithm for mutual exclusivity detection (FSME). FSME initially searches all paired gene sets to find potential seed genes, then expands the seed genes to form gene sets, and finally tests the significance of the mutual exclusivity of the gene sets.

Results: For the breast cancer data, the SME method identified a stratified mutually exclusive gene set: CDH1, TP53, GATA3, MAP3K1, FAT1. This gene set accounts for 65.73% of the variation in all samples. When using FSME across different subtypes, the results in the Luminal-A subtype were highly consistent with the overall results, indicating that the predominant subtype has a strong influence on the overall results. In the Basal subtype, the mutually exclusive gene set DST, TP53, and CFAP47 had a total variation rate of 92.86%, with TP53 being a tumor suppressor gene not specific to breast cancer, while DST and CFAP47 might play important roles in the progression of breast cancer.

Conclusion: The SME and FSME methods proposed in this study effectively identify mutually exclusive genes in breast cancer, ensuring high precision and low false discovery rates. By analyzing different subtypes, important mutually exclusive gene sets were identified, particularly in the Basal subtype where the potential driver gene DST was discovered. These findings provide new insights for the precision treatment of breast cancer and have significant clinical implications.

**Session 24CHI88: Recent Developments in Complex Time-to-Event Data Analysis**

**Transformation Models with Informative Partly Interval-Censored Data**

♦*Jingjing Jiang1, Chunjie Wang1, Deng Pan2* and *Xinyuan Song3*

1Changchun University of Technology， 2Huazhong University of Science and Technology， 3The Chinese University of Hong Kong

Partly interval censoring is frequently encountered in clinical trials when the failure time of an event is observed exactly for some subjects but is only known to fall within an observed interval for others. Although this kind of censoring has drawn recent attention in survival analysis, available methods typically assume that the observed interval is independent of the failure time and that all potential predictors can be fully observable. However, the above assumptions maynot be valid in practice. This paper considers a new joint modeling approach to simultaneously model the failure and observation times and correlate these two stochastic processes through shared latent factors. The proposed model comprises a transformation model for the failure time of interest, a proportional hazards model for the length of censoring interval, and a factor analysis model for characterization of the latent factors. A multi-stage data augmentation procedure is introduced to tackle the challenges posed by the complex model and data structure. A Bayesian approach coupled with monotone spline approximation and Markov chain Monte Carlo techniquesis developed to estimate the unknown parameters and nonparametric functions. The satisfactory performance of the proposed method is demonstrated through simulations, and it is then applied to a Framingham Heart study.

**Additive Hazards Model with Time-Varying Coefficients and Imaging Predictors**

♦*Qi Yang1*

1Shandong University

This study considers a time-varying coefficient additive hazards model with latent variables to examine potential observed and latent risk factors for survival of interest. The model consists of two parts: confirmatory factor analysis to measure each latent factor through multiple observable variables and a varying coefficient additive hazards model to examine the time-varying effects of the observed and latent risk factors on the hazard function. A hybrid estimation procedure that combines the expectation-maximum algorithm and corrected estimating equation method is developed to estimate the unknown parameters and nonparametric cumulative hazard functions. The consistency and asymptotic normality of the proposed estimators are established, and the pointwise confidence intervals and general confidence bands of the nonparametric functions are constructed accordingly. A satisfactory performance of the proposed method is demonstrated

through simulation studies. An application to a study of chronic kidney disease for Chinese type 2 diabetes patients is presented.

**Accelerated Hazards Model Under Length-Biased Sampling**

*Cong Liu1* and♦*Haijin He2*

1Shunfeng, 2Shenzhen University

The accelerated hazards model is often used in survival analysis to study differences in treatment effects of biomedical trials. The model specifies that covariates influence the hazard function through a time-scale change. The accelerated hazards model relaxes the proportional hazards assumption of traditional models, and thus can be applied when the treatment group and the control group have equal hazard at the beginning of study. Length-biased data are not randomly selected from the population, but with probability proportional to their lengths. Length-biased sampling is commonly encountered in randomized experiments or observational studies. Ignoring the feature of length-biased data may lead to erroneous or misleading analysis. We propose an accelerated hazards model under length-biased sampling. The proposed model owns the special features and advantages of the common accelerated hazards model, and properly handles length-biased data. Based on the generating scheme of length-biased data, we propose a bias-weight-adjusted estimating equation method to obtain the regression parameter estimates. We prove the asymptotic properties of the proposed estimators. Simulations are conducted to evaluate the finite sample performance of the proposed method, and real data analysis are provided to demonstrate the utility of the proposed methodology.

**Regression Analysis of Group-Tested Current Status Data**

♦*Shuwei Li*

Group testing is an effective way to reduce the time and cost associated with conducting large-scale screening for infectious diseases. Benefits are realized through testing pools formed by combining specimens, such as blood or urine, from different individuals. In some studies, individuals are assessed only once and a time-to-event endpoint is recorded, for example, the time until infection. Combining group testing with this type of endpoint results in group-tested current status data (Petito & Jewell, 2016). To analyse these complex data, we propose methods that estimate a proportional hazard regression model based on test outcomes from measuring the pools. A sieve maximum likelihood estimation approach is developed that approximates the cumulative baseline hazard function with a piecewise constant function. To identify the sieve estimator, a computationally efficient expectation-maximization algorithm is derived by using data augmentation. Asymptotic properties of both the parametric and nonparametric components of the sieve estimator are thenestablished by applying modern empirical process theory. Numerical results from simulation studies show that our proposed method performs nominally and has advantages over the corresponding estimation method based on individual testing results. We illustrate our work by analysing a chlamydia dataset collected by the State Hygienic Laboratory at the University of Iowa.

**Session 24CHI125: Statistics in Biosciences**

**A Semiparametric Cox-Aalen Transformation Model with Censored Data**

*Xi Ning1, Yinghao Pan2,* ♦*Yanqing Sun2* and *Peter Gilbert3*

1Colby College， 2University of North Carolina at Charlotte， 3Department of Biostatistics, University of Washington and Fred Hutchinson Cancer Center

We propose a broad class of so-called Cox-Aalen transformation models that incorporate both multiplicative and additive covariate effects on the baseline hazard function within a transformation. The proposed models provide a highly flexible and versatile class of semiparametric models that include the transformation models and the Cox-Aalen model as special cases. Specifically, it extends the transformation models by allowing potentially time-dependent covariates to work additively on the baseline hazard and extends the Cox-Aalen model through a predetermined transformation function. We propose an estimating equation approach and devise an Expectation-Solving (ES) algorithm that involves fast and robust calculations. The resulting estimator is shown to be consistent and asymptotically normal via modern empirical process techniques. The ES algorithm yields a computationally simple method for estimating the variance of both parametric and nonparametric estimators. Finally, we demonstrate the performance of our procedures through extensive simulation studies and applications in two randomized, placebo-controlled HIV prevention efficacy trials. The data example shows the utility of the proposed Cox-Aalen transformation models in enhancing statistical power for discovering covariate effects.

**Risk Predictions Using Data with Complex Censoring Mechanism**

♦*Haoxuan (Charlie) Zhou1, Xiaoqiong Joan Hu1, Yan Yuan2* and *Yi Xiong3*

1Simon Fraser University， 2University of Alberta， 3University at Buffalo

We aim to develop a tool for risk predictions with a combination of doubly-censored and interval-censored event times. This project is partly motivated by a research project from the Childhood Cancer Survivor Study to predict the primary ovarian insufficiency (POI) risk in female childhood cancer survivors. The study data include self-reported demographic information and 5-year cancer treatment history. The study subjects’ ages at POI are recorded or known within certain time interval if the POI was experienced after the cancer diagnosis. That results in the available data on age at POI with an unconventional structure. We formulate the study data into a combination of double-censored andinterval-censored event times. Natural and easy-to-implement extensions of two commonly-used approaches, (i.e. Cox proportional hazards model and age-specific logistic regression) to risk prediction are proposed in the settings. We examine the proposed approaches via simulation. The POI data from the Childhood Cancer Survivor Study are used throughout the talk to motivate and illustrate the proposed approaches. This is joint work with Joan Hu, Yi Xiong, and Yan Yuan.

**Use of Real-World Electronic Health Records (EHR) Data to Identify Potential Adverse Effects of COVID-19 Vaccines**

♦*Hulin Wu1,*

*1*University of Texas Health Science Center at Houston

Real-world electronic health record (EHR) data can be used to identify rare adverse events or side effects of medications or treatments because of its large sample size. In this study, we use a large nationwide EHR COVID-19 database to identify potential adverse effects of COVID-19 vaccines by comparing to historical EHR data of influenza vaccination. To compare the proportion of adverse events in a pre-specified time interval after vaccination, the censored or incomplete data should be considered. We propose an inverse probability of censoring weight (IPCW) adjusted hypothesis test to deal with the censored EHR data problem. The asymptotic consistent estimator of the event proportion is proposed and the corresponding asymptotic distribution is derived under the data censoring. The asymptotic hypothesis test for comparing the event proportion between the two groups under censoring is established. The simulation studies are conducted to demonstrate the validity of the proposed IPCW-adjusted estimate and test statistic. In particular, we show that the proposed IPCW-adjusted estimate of the event proportion can remove the estimation bias and the type I error of IPCW-adjusted test can be controlled well at the nominal level when the censoring rate increases. The proposed IPCW-adjusted test is applied to COVID-19 EHR data to identify potential adverse effects of COVID-19 vaccines.

**The Analysis of Gene Expression Using Statistical Learning Approaches and Research Teaching Innovation**

*Xiyuan Liu*1, ♦*Di Gao*2 and *Gang Shen*3.

1Louisiana Tech University, 2Sam Houston State University，3North Dakota State University.

The classification problem of gene expression level, specifically gene expression analysis, is a major research area in biostatistics. There are several classical methods to solve the classification problem. For example, hypothesis testing with the Benjamini-Hochberg method, Logistic regression, Hidden Markov Chain (HMM) model, and Support Vector Machine (SVM). These classical methods can be applied and perform well with some strict assumptions. One of such is the observations in the data should beindependent. However, in gene analysis, these assumptions are violated. The Conditional Random Field (CRF) was introduced to solve the problem. This presentation will discuss some classical methods and the CRF model. Finally, the Least Absolute Selection and Shrinkage Operator (LASSO) penalty, a dimensional reduction method, is introduced to improve the CRF model.

**Session 24CHI111: Statistical Considerations for Rare Disease and Pediatric Clinical Trials with Challenges in Study Design**

**Regional Sample Size in MRCT–Statistical Considerations**

♦*Weiya Xu1* and *Hao Zhang1* .

1Sanofi

MRCT is attractive because it allows examination of treatment effect based on data from all regions and facilitate simultaneous global development of a drug and reduce the number of clinical studies conducted separately in each region. There is no uniformly accepted or optimal approach of sample size allocation to regions in an MRCT. Preservation of effect (PMDA Method 1 and Method 2) is mainly used in sample size allocation to China subpopulation. This presentation will introduce some deeper statistical considerations with cases.

**Estimand Consideration for Recurrent Event**

♦*Wenruo Hu*1 .

1Sanofi

Recurrent events refer to the repeated occurrence of the same type of event over time for the same participant. It increasingly draws attention in recent clinical trials as it could describe full burden of disease and maybe more relevant to evaluate treatment effect. Moreover, it could increase statistical power comparing to endpoints focus on the first event or an assessment at fix timepoint. This is one of the appealing aspects in statistical consideration especially in rare disease setting. Currently, recurrent events endpoint is used as primary endpoint in many rare diseases, for example, annualized relapse rate (ARR) and MRI lesion count in multiple sclerosis and annualize bleeding rate (ABR) in hemophilia. Estimand concept was highlighted in ICH E9 (R1) addendum in 2019. It is required topic in clinical trials afterwards. In this presentation, we would like to discuss statistical consideration of estimand in recurrent event endpoint to facilitate choice of estimand strategy in clinical trial.

**Statistical Considerations on Leveraging External Data in Clinical Trials**

♦*Grace Lin1* .

1Sanofi

There is an increasing regulatory interest in synthesizing evidence from current (randomized) clinical trials with other data sources, to better understand the safety and efficacy of new drugs, especially in rare disease and pediatric clinical trials. Leveraging external data has been an appealing strategy to supplement the analysis of trials and improve decision-making. However, the combination of trial and external data introduces several sources of potential bias that need to be attenuated. Several methods have been established and developed to borrow external data, among which the propensity score methods and Bayesian dynamic borrowing framework play essential roles. Bayesian dynamic borrowing which can discount the historical data based on the outcome similarity of the two populations, a considerable population difference may still lead to a moderate bias. Hence, a robust adjustment for the population difference using approaches such as propensity score, can make the borrowing more efficient and robust. In this presentation, we will evaluate the effect of external borrowing using integrated approach that combines propensity score and Bayesian dynamic borrowing methods.

**Statistical Considerations in Defining Causal Estimands of Single-Arm Trials**

♦*Liqi Feng* and *Hao Zhang*.

The design of single arm trials are widely adopted in pediatric studies and rare diseases studies. It is not intuitive to define an estimand in single arm trials while causal conclusions may needed to be made. We will show some of the methods used in defining a causal estimand in single-arm trials and present our thinking on the statistical considerations in such study design.

**Session 24CHI9: Advancements in Statistical Testing, Estimation, and Modeling Across Diverse Data Structures**

**Estimation and Inference for Fixed Center Effects on Panel Count Data**

♦*Weiwei Wang1, Yijun Wang1* and *Xiaobing Zhao2*

1Zhejiang Gongshang University，2Zhejiang University of Finance and Economics

In familial or multi-center studies, comparisons of outcomes across different centers are of interest. An extensive body of statistical models has been developed for various outcomes. However, most existing models apply only to simple data types. In this article, a fixed center effect proportional mean model is suggested to quantify center effects with respect to panel count data. When the number of centers is large, the traditional estimation methods that treat thesecenter effects as categorical variables have many parameters to be estimated and thus may not be feasible to implement. In order to avoid including so many unknown variables, a new estimation procedure is proposed, where the center effects can be easily estimated by the center-specific ratio of observed to expected cumulative numbers of panel count data. The dimension of the estimated parameter space of the proposed procedure is only dependent on the number of covariates, and it is computationally more efficient. Given some regularity conditions, the asymptotic properties of the proposed estimators are established. Extensive simulation studies are conducted to assess the finite-sample properties of the proposed estimators. Finally, the proposed method is applied to a real dataset from the China Health and Nutrition Study.

**Two-Sample Tests Based on Data Depth**

*XIaoping Shi1, Yue Zhang1* and♦*Chifeng Shen2*

1University of British Columbia Okanagan，2York University

This talk focuses on the homogeneity test that evaluates whether two multivariate samples come from the same distribution. The problem arises naturally in various applications, and many methods are available in the literature. Based on data depth, several tests have been proposed for this problem, but they may not be very powerful. In light of the recent development of data depth as an important measure of quality assurance, two new test statistics are proposed for the multivariate two-sample homogeneity test. The proposed statistics, namely the maximum statistics and rotated ellipse statistics, exhibit a common chi-squared asymptotic null distribution. Furthermore, the generalization of the proposed tests into the multivariate multi-sample situation is also discussed. Simulation studies demonstrate the superior performance of the proposed tests, with a real data example illustrating their practical application.

**Consistent Semiparametric Tests for Lorenz Dominance Based on the Density Ratio Model**

*Wenchen Liao*1, *Weiqi Yang*1, *Yukun Liu*2 and ♦*Weiwei Zhuang*1.

1University of Science and Technology of China, 2East China Normal University

Lorenz dominance is a fundamental tool for judging whether wealth disparity expands further in one country or region than another. The density ratio model, as a kind of semiparametric model, has good performance in analyzing homogeneous data. This paper proposes a new estimator for the Lorenz curve under that model. In addition, we construct a new statistic to test Lorenz dominance, and investigate its asymptotic properties. We also provide a valid bootstrap method for hypothesis testing. Simulations show that our proposed estimator substantially improves the estimation efficiency and power of the test. Finally, we apply our method to compare the individual wages and salaries of the USA in different years and analyze the changes in county-level $CO\_2$ emissions of China in 2007 and 2017.

**Efficient Auxiliary Information Synthesis for Cure Rate Model**

*Jie Ding*, *Jialiang Li*, ♦*Xiaoguang Wang*

We propose a new auxiliary information synthesis method to utilize subgroup survival information at multiple time points under the semi-parametric mixture cure rate model. After summarizing the auxiliary information via estimating equations, a control variate technique is adopted to reduce the variance efficiently, together with a test statistic to check the homogeneity assumption. Revision using penalization is further considered to adaptively accommodate potential population heterogeneity. Our methods can be adjusted when the uncertainty is not negligible. We establish asymptotic properties of our proposed estimators, and

demonstrate their practical performances through extensive simulations and an invasive breast cancer study.

**Session 24CHI9: Advancements in Statistical Testing, Estimation, and Modeling Across Diverse Data Structures**

**Estimation and Inference for Fixed Center Effects on Panel Count Data**

♦ *Weiwei Wang1, Yijun Wang1* and *Xiaobing Zhao2*

1Zhejiang Gongshang University，2Zhejiang University of Finance and Economics

In familial or multi-center studies, comparisons of outcomes across different centers are of interest. An extensive body of statistical models has been developed for various outcomes. However, most existing models apply only to simple data types. In this article, a fixed center effect proportional mean model is suggested to quantify center effects with respect to panel count data. When the number of centers is large, the traditional estimation methods that treat these center effects as categorical variables have many parameters to be estimated and thus may not be feasible to implement. In order to avoid including so many unknown variables, a new estimation procedure is proposed, where the center effects can be easily estimated by the center-specific ratio of observed to expected cumulative numbers of panel count data. The dimension of the estimated parameter space of the proposed procedure is only dependent on the number of covariates, and it is computationally more efficient. Given some regularity conditions, the asymptotic properties of the proposed estimators are established. Extensive simulation studies are conducted to assess the finite-sample properties of the proposed estimators. Finally, the proposed method is applied to a real dataset from the China Health and Nutrition Study.

**Two-Sample Tests Based on Data Depth**

*Xiaoping Shi1, Yue Zhang1* and♦*Chifeng Shen2*

1University of British Columbia Okanagan，2York University

This talk focuses on the homogeneity test that evaluates whether two multivariate samples come from the same distribution. The problem arises naturally in various applications, and many methods are available in the literature. Based on data depth, several tests have been proposed for this problem, but they may not be very powerful. In light of the recent development of data depth as an important measure of quality assurance, two new test statistics are proposed for the multivariate two-sample homogeneity test. The proposed statistics, namely the maximum statistics and rotated ellipse statistics, exhibit a common chi-squared asymptotic null distribution. Furthermore, the generalization of the proposed tests into the multivariate multi-sample situation is also discussed. Simulation studies demonstrate the superior performance of the proposed tests, with a real data example illustrating their practical application.

**Consistent Semiparametric Tests for Lorenz Dominance Based on the Density Ratio Model**

*Wenchen Liao*1, *Weiqi Yang*1, *Yukun Liu*2 and ♦*Weiwei Zhuang*1

1University of Science and Technology of China, 2East China Normal University

Lorenz dominance is a fundamental tool for judging whether wealth disparity expands further in one country or region than another. The density ratio model, as a kind of semiparametric model, has good performance in analyzing homogeneous data. This paper proposes a new estimator for the Lorenz curve under that model. In addition, we construct a new statistic to test Lorenz dominance, and investigate its asymptotic properties. We also provide a valid bootstrap method for hypothesis testing. Simulations show that our proposed estimator substantially improves the estimation efficiency and power of the test. Finally, we apply our method to compare the individual wages and salaries of the USA in different years and analyze the changes in county-level $CO\_2$ emissions of China in 2007 and 2017.

**Efficient Auxiliary Information Synthesis for Cure Rate Model**

*Jie Ding*, *Jialiang Li*, ♦*Xiaoguang Wang*

We propose a new auxiliary information synthesis method to utilize subgroup survival information at multiple time points under the semi-parametric mixture cure rate model. After summarizing the auxiliary information via estimating equations, a control variate technique is adopted to reduce the variance efficiently, together with a test statistic to check the homogeneity assumption. Revision using penalization is further considered to adaptively accommodate potential population heterogeneity. Our methods can be adjusted when the uncertainty is not negligible. We establish asymptotic properties of our proposed estimators, and

demonstrate their practical performances through extensive simulations and an invasive breast cancer study.

**Session 24CHI25: Complex Statistical Inference for Large-Scale Datasets**

**A Communication Efficient Boosting Method for Distributed Spectral Clustering**

♦*Yingqiu Zhu1 and Danyang Huang2*

1University of International Business and Economics， 2Renmin University of China

Spectral clustering is one of the most popular clustering techniques in statistical inference. For large-scale datasets, spectral clustering is typically implemented through distributed computing. However, existing distributed implementations face two major challenges. First, the clustering performance is negatively affected by distributed computing since the topological structure of all objects has to be divided into distributed parts. Second, the communication among computers within a distributed system results in high communication costs. To address these issues, we propose a communication-efficient algorithm for distributed spectral clustering. Our motivation stems from a theoretical comparison between the conventional spectral clustering algorithm, which operates on the entire dataset, and the local spectral clustering, which is performed on a subsample using a single computer. We identify the critical factor that leads to the difference between the performances of global spectral clustering and local spectral clustering. Based on the findings, we propose a novel approach that iteratively aggregates the intermediate results generated by local spectral clustering. In this process, only low-dimensional vectors are exchanged between computers. The results of simulations and real data analysis demonstrate that the proposed method apparently enhances the performance of distributed spectral clustering with low communication costs.

**Latent Space Link Prediction Model for Patent Citation**

♦*Wei Hu and Ning Yang.*

Patent citation relationships constitute a citation network, and the predictability of edges in the network is a frontier research issue in complex networks. This article explores the predictability of patent citation relationships, and provides methodological and technical support for technology diffusion. We proposes a new link prediction model, namely a latent space link prediction model of patent citation relationships. The model measures the impact of technical text content, local network structure, individual heterogeneity and other factors on link predictability. The parameter estimation algorithm, citation relationship prediction algorithm, and prediction result evaluation index system of the model are proposed. Finally, empirical research is conducted in the field of quantum sensing. Research results demonstrate that patent citation relationships exhibit the characteristics of technical content correlation, time variability, local network structure similarity, and individual heterogeneity. Moreover, the patent citation has a positive U-shaped relationship with the time interval between patents. The model proposed in this article could improve the prediction performance of patent citation, and provide support for patent review, technology diffusion paths and patent value assessment.

**Pairwise Maximum Likelihood for Multi-Class Logistic Regression Model with Multiple Rare Classes**

♦*Xuetong Li1, Danyang Huang2 and Hansheng Wang1.*

1Peking University， 2Renmin University of China

We study in this work the problem of multi-class logistic regression with one major class and multiple rare classes. It is motivated by a real application in TikTok live stream data. The model is inspired by the two-class logistic regression model of Wang (2020) but with surprising theoretical findings, which in turn motivate new estimation methods with excellent statistical and computational efficiency. Specifically, rigorous theoretical analysis suggests that the resulting maximum likelihood estimators of different rare classes should be asymptotically independent. Hence, instead of optimizing the joint log-likelihood function with computational challenge in multi-class problem, we consider to solve multiple pairwise two-class logistic regression problems, which are computationally much easier and can be conducted in a fully parallel way. We show rigorously that the resulting estimator could be as asymptotically efficient as the global maximum likelihood estimator under appropriate regularity conditions. To further reduce the computation cost, a subsample-based pairwise likelihood estimator is developed by down-sampling the major class. The resulting estimator is also asymptotically efficient if the sampling probability is appropriately specified.

Extensive simulation studies are presented to support our theoretical findings and a TikTok live stream dataset is analyzed for illustration purpose.

**Subsampling Strategy of AIC-based Model Averaging for Generalized Linear Models with Massive datasets**

♦*Jun Yu*1, *HaiYing Wang*2 *and Mingyao Ai*3.

1Beijing Institute of Technology, 2University of Connecticut and 3Peking University.

Subsampling is an effective approach to dealing with computational challenges with massive data. However, existing subsampling methods do not take into account model uncertainty. In this paper, we investigate the subsampling technique for the Akaike information criterion (AIC) and extend the subsampling method to the smoothed AIC model averaging framework in the context of generalized linear models. By correcting the asymptotic bias of the maximized subsample objective function that is used to approximate the Kullback–Leibler divergence, we derive the form of the AIC based on the subsampled data. We then provide a subsampling strategy for the smoothed AIC model averaging estimator and study the corresponding asymptotic properties of the loss and the resulting estimator. A practically implementable algorithm is developed and its performance is evaluated through numerical experiments on both simulated and real datasets.

**Session 24CHI32: Frontiers in Multiple Hypothesis Testing**

1. **values, Multiple Testing and Beyond**

♦*Xianyang Zhang1 and Guanxun Li1*

1Texas A&M University

We discover a connection between the Benjamini-Hochberg (BH) procedure and the recently proposed e-BH procedure [Wang and Ramdas, 2022] with a suitably defined set of e-values. This insight extends to a generalized version of the BH procedure and the model-free multiple testing procedure in Barber and Candes [2015] (BC) with a general form of rejection rules. The connection provides an effective way of developing new multiple testing procedures by aggregating or assembling e-values resulting from the BH and BC procedures and their use in different subsets of the data. In particular, we propose new multiple testing methodologies in three applications, including a hybrid approach that integrates the BH and BC procedures, a multiple testing procedure aimed at ensuring a new notion of fairness by controlling both the group-wise and overall false discovery rates (FDR), and a structure adaptive multiple testing procedure that can incorporate external covariate information to boost detection power. One notable feature of the proposed methods is that we use a data-dependent approach for assigning weights to e-values, significantly enhancing the efficiency of the resulting e-BH procedure. The construction of the weights is non-trivial and is motivated by the leave-one-out analysis for the BH and BC procedures. In theory, we prove that the proposed e-BH procedures with data-dependent weights in the three applications ensure finite sample FDR control. Furthermore, we demonstrate the efficiency of the proposed methods through numerical studies in the three applications.

**A Constructive Approach to Selective Risk Control**

♦*Zijun Gao1, Wenjie Hu2 and Qingyuan Zhao3.*

1University of Southern California，2Peking University， 3University of Cambridge

Many modern applications require the use of data to both select the statistical tasks and make valid inference after selection. In this article, we provide a unifying approach to control for a class of selective risks. Our method is motivated by a reformulation of the celebrated Benjamini-Hochberg (BH) procedure for multiple hypothesis testing as the iterative limit of the Benjamini-Yekutieli (BY) procedure for constructing post-selection confidence intervals. Building on this observation, we propose a constructive approach to control extra-selection risk (selection made after decision) by iterating decision strategies that control the post-selection risk (decision made after selection), and show that many previous methods and results are special cases of this general framework. Our development leads to two surprising results about the BH procedure: (1) in the context of one-sided location testing, the BH procedure not only controls the false discovery rate at the null but also at other locations for free; (2) in the context of permutation tests, the BH procedure with exact permutation p-values can be well approximated by a procedure which only requires a total number of permutations that is almost linear in the total number of hypotheses.

**Integrative Conformal p-values for Out-of-distribution Testing with Labeled Outliers**

*Ziyi Liang1, Matteo Sesia1 and* ♦*Wenguang Sun2*

1University of Southern California，2Zhejiang University

This talk presents novel conformal inference methods for out-of-distribution testing that leverage side information from labeled outliers, which are commonly underutilized or even discarded by conventional conformal p-values. Blending inductive and transductive conformal inference strategies in a principled way, our methods are computationally efficient and can automatically take advantage of the most powerful model from a collection of one-class and binary classifiers. Then, we study how to control the false discovery rate in multiple testing with a conditional calibration strategy. Simulations with synthetic and real data show the proposed integrative conformal p-values outperforms existing methods. This is the joint work with Ziyi Liang and Matteo Sesia.

**Asymptotic Uncertainty of False Discovery Proportion**

*Meng Mei,* ♦*Tao Yu1 and Yuan Jiang2.*

1Department of Statistics and Data Science, National University of Singapore，2Department of Statistics, Oregon State University

Multiple testing has been a popular topic in statistical research. Although vast works have been done, controlling the false discoveries remains a challenging task when the corresponding test statistics are dependent. Various methods have been proposed to estimate the false discovery proportion (FDP) under arbitrary dependence among the test statistics. One of the main ideas is to reduce arbitrary dependence to weak dependence and then to establish theoretically the strong consistency of the FDP and false discovery rate (FDR) under weak dependence. As a consequence, FDPs share the same asymptotic limit in the framework of weak dependence. We observe that the asymptotic variance of the FDP, however, may rely heavily on the dependence structure of the corresponding test statistics even when they are only weakly dependent; and it is of great practical value to quantify this variability, as it can serve as an indicator of the quality of the FDP estimate from the given data. As far as we are aware, the research on this respect is still limited in the literature. In this paper, we first derive the asymptotic expansion of FDP under mild regularity conditions and then examine how the asymptotic variance of FDP varies under different dependence structures both theoretically and numerically. With the observations in this study, we recommend that in a multiple testing performed by an FDP procedure, we may report both the mean and the variance estimates of FDP to enrich the study outcome.

**Session 24CHI43: Machine Learning Methods for Multi-Omics Data Analysis**

**Integrated Analysis of Imaging and RNA-Seq to Characterize Smoking-Related Lung Disease Phenotypes**

♦*Fenghai Duan1.*

1Brown University School of Public Health

The goal of this study was to uncover biological distinctions across various smoking-related phenotypes by integrating clinical, imaging, and bronchial epithelial RNA-Seq data. Utilizing K-means clustering, we grouped participants based on computed tomography (CT) imaging characteristics and analyzed their clinical phenotypes. Our analysis was augmented by examining bronchial epithelial gene expression through methods such as differential gene expression analysis, over-representation analysis, gene set enrichment analysis, and gene set variation analysis.

Three distinct clusters were identified: preserved, interstitial predominant, and emphysema predominant. Compared to the preserved cluster, the interstitial and emphysema clusters demonstrated poorer lung function, reduced exercise capacity, and lower quality of life. Additionally, they exhibited more rapid declines in exercise capacity and lung function over time, increased progression of emphysema, a higher incidence of acute respiratory disease events, and greater mortality rates. The emphysema cluster, in particular, exhibited the most severe clinical outcomes, followed by the interstitial cluster. Our transcriptomic analysis indicated that the more severe disease stages were associated with increased inflammatory responses, especially through the TNF-α pathway, while less severe stages were characterized by an upregulation of T-cell-related genes.

By employing quantitative CT imaging, we identified three subgroups of disease among individuals with a history of smoking. The differences in airway gene expression among these groups suggest a link between the severity of the clinical phenotype and elevated inflammation, possibly mediated by the TNF-α pathway.

**Can We Make Learning Machines to Discover New Knowledge?**

♦*Xuegong Zhang1*.

1Tsinghua University

Many people are talking about AI for science nowadays, especially AI and Machine Learning for Life Sciences. Life is a complex information system. Inferring the underlying mechanisms of life from a large amount of data is a great challenge for both human intelligence and artificial intelligence. Since the birth of bioinformatics, machine learning has played an important role in the exploration of life sciences, playing an important role in the analysis of various high-throughput biological data. But most of these roles are to fulfil specific tasks designed by humans. Is it possible for machine learning and artificial intelligence to discover complex patterns behind biological processes? We proposed the concept of "discovery machine learning" and made preliminary attempts to construct cell differentiation lineages using early human embryonic single-cell data, exploring whether AI can make biological discoveries with least involvement of human knowledge or guide. We also attempted to apply similar ideas to the analysis of early lung cancer imaging data and explored the use of large foundation models to learn gene expression language in single-cell transcriptome data. This talk will share some of our exploration experiences and reflections on future development.

**Powerful and Accurate Detection of Temporal Gene Expression Patterns from Multi-Sample Multi-Stage Single-cell Transcriptomics Data with TDEseq**

♦*Shiquan Sun1*.

*1*Xi'an Jiaotong University

We present a non-parametric statistical method called TDEseq that takes full advantage of smoothing splines basis functions to account for the dependence of multiple time points in scRNA-seq studies, and uses hierarchical structure linear additive mixed models to model the correlated cells within an individual. As a result, TDEseq demonstrates powerful performance in identifying four potential temporal expression patterns within a specific cell type. Extensive simulation studies and the analysis of four published scRNA-seq datasets show that TDEseq can produce well-calibrated p-values and up to 20% power gain over the existing methods for detecting temporal gene expression patterns.

**Intelligent Spatial Transcriptomics: Methods and Applications**

♦*Shihua Zhang1*.

*1*Academy of Mathematics and Systems Science, CAS

Technological advances in spatial transcriptomics are critical for a better understanding of the structures and functions of tissues in biological research. The combination of intelligent or statistical algorithms and spatial transcriptomics has emerged to pave the way for deciphering tissue architecture. We have made great efforts to advance intelligent spatial transcriptomics and developed a group of STA- tools. For example, we first developed a graph attention auto-encoder tool STAGATE to identify spatial domains by learning low-dimensional latent embeddings via integrating spatial information and gene expression profiles. Second, we introduced STAligner for integrating and aligning ST datasets across different conditions, technologies, and developmental stages to enable spatially-aware data integration, simultaneous spatial domain identification, and downstream comparative analysis. Third, we designed STAMarker for identifying spatially domain-specific variable genes with saliency maps in deep learning. Fourth, we developed a spatial location-supervised auto-encoder generator STAGE for generating high-density spatial transcriptomics. Fifth, we developed STASCAN for deciphering fine-resolution cell-distribution maps in spatial transcriptomics.

**Session 24CHI14: Advances in Theories and Methods for Dependent Data Analysis**

**MSB: A Unified Bootstrap for $L^infty$-norm Statistics of High-Dimensional Time Series**

*Ruru Ma*1 *and* ♦*Shibin Zhang*1

1Shanghai Normal University

$L^infty$-norm statistics are commonly encountered in high-dimensional data analysis. When constructed from high-dimensional time series, their asymptotic distributions often perform poorly in finite-sample inference. Various multiplier bootstraps, tailored differently for each $L^infty$-norm statistic, are typically employed to approximate the sampling distributions. In this talk, a unified multiplier bootstrap approach, called the multiplier subsample bootstrap (MSB), is proposed to approximate the sampling distribution of $L^infty$-norm statistics constructed from high-dimensional time series. Assuming that the cumulative distribution function (cdf) of an $L^infty$-norm statistic converges uniformly to the cdf of a Gaussian vector's $L^infty$-norm, we establish the asymptotic validity of the MSB estimator. A simulation study demonstrates that the proposed MSB achieves good performance.

**Self-weighted Quantile Estimation of Ornstein-Uhlenbeck Process with Jumps**

♦*Yuping Song*1, *Chunchun Cai*1 *and Min Zhu*1

1Shanghai Normal University

The self-weighted quantile regression estimator for the drift parameter in diffusion models with jumps is proposed. The asymptotic normality of the underlying estimator is obtained. Moreover, the better finite-sample properties are verified through the Monte-Carlo simulation study and

empirical analysis.

**On Log-Concave-Tailed Chaos and the Restricted Isometry Property**

*Guozheng Dai*1, *Zhonggen Su*1, *Vladimir Ulyanov*2 *and* ♦*Hanchao Wang*3.

1Zhejiang University， 2Moscow State University， 3Shandong University

We develop a novel decoupling inequality for nonhomogeneous non-Gaussian chaoses of order 2. Having this decoupling inequality, we get two types of the $p$-th moment bounds for the suprema of a log-concave-tailed chaos process based on some chaining techniques and the generalized majorizing measure theorems for canonical processes. As applications, we prove the restricted isometry property of partial random circulant matrices and time-frequency structured random matrices induced by standard $\alpha$-subexponential vectors, $1\le \alpha\le 2$, which extends the previously known results for the sub-Gaussian case.

**A Variation-Ratio Test for Volatility Jumps Using Noisy High Frequency Data**

♦*Guangying Liu, Yang Li and Zhiyuan Zhang*

This paper proposes a novel variation-ratio test for the presence of volatility jumps using high frequency data with microstructure noise. Under the null hypothesis

that the volatility process is a continuous semimartingale, the test statistic is asymptotically normal. Under the alternative hypothesis that the volatility process jumps, the test statistic diverges at a rate (n^(1/4-l) for arbitrarily small l > 0) faster than the best rate (close to n^(1/8) in the semimartingale setup) available in the literature. Simulation results corroborate our theoretical ndings. Empirical results show that our test fails to reject the null hypothesis for most of the ninety US stocks studied.

**Session 24CHI44: Machine Learning Methods for Survival Data**

**Bayesian Tree-based Heterogeneous Mediation Analysis with a Time-to-Event Outcome**

*Rongqian Sun*1 ，♦*Xinyuan Song*1

1The Chinese University of Hong Kong

Mediation analysis aims at quantifying and explaining the underlying causal mechanism between an exposure and an outcome of interest. In the context of survival analysis, mediation models have been widely used to achieve causal interpretation for the direct and indirect effects on the survival of interest. Although heterogeneity in treatment effect is drawing increasing attention in biomedical studies, none of the existing methods have accommodated the presence of heterogeneous causal pathways pointing to a time-to-event outcome. This study considers a heterogeneous mediation analysis for survival data based on a Bayesian tree-based Cox proportional hazards model with shared topologies. Under the potential outcomes framework, individual-specific conditional direct and indirect effects are derived on the scale of the logarithm of hazards, survival probability, and restricted mean survival time. A Bayesian approach with efficient sampling strategies is developed to estimate the conditional causal effects through the Monte Carlo implementation of the mediation formula. Simulation studies show the satisfactory performance of the proposed method. The proposed model is then applied to an HIV dataset extracted from the ACTG175 study to demonstrate its usage in detecting heterogeneous causal pathways.

**Gradient Boosting Methods for Survival Data**

*Peizhi Li*1, ♦*Yingwei Peng*2 and *Jianing Zheng*1

1Dongbei University of Finance and Economics， 2Queen's University

Gradient boosting methods become popular in recent years to analyze right-censored survival data where Cox's proportional hazards model is the widely used statistical model. However, there are very limited studies on the differences between the two approaches for right-censored survival data. We first compared two boosting methods with Cox's proportional hazards model in this paper: one is the gradient boosting decision tree and the other is gradient boosting with component-wise linear models. The differences between the two boosting methods are studied numerically. The results show that the boosting methods outperform Cox's proportional hazards model in both the relative and absolute risk estimation in the proportional hazards model except when Cox's proportional hazards model is fully specified with nonlinear and interaction covariates effects. It indicates that the boosting methods, particularly the gradient boosting decision tree, is a very competitive method for right-censored survival data if complicated covariate effects may exist but are unknown to the investigator. We further proposed a gradient boosting decision tree based method to estimate the mixture cure model for survival data with a cured fraction. The new method inherits the features of the GBDT and provides more accurate estimates of the cure probability and the relative risk for uncured subjects than existing methods when there are no a priori assumptions on the forms of complex covariate effects in the model.

**Polya Trees for Survival Data**

♦*Liqun Diao*1 and *Yixing Zhao*1

1University of Waterloo

Polya trees are commonly used as priors in nonparametric Bayesian analysis. This presentation will discuss approaches for utilizing Polya trees to characterize the distribution of survival data, which may be subjected to different forms of censoring such as right-censoring or interval-censoring. The discussion will cover different aspects of Polya trees, including partitions, prior strength, and choices of prior distributions. Comparisons of the proposed methods to existing approaches for estimating survival probabilities are provided in both simulated settings and through applications to real datasets. It is shown that the proposed methods either improve upon or remain competitive with existing nonparametric estimation methods.

**DeepSurv Landmarking: A Deep Learning Approach for Dynamic Survival Analysis with Longitudinal Data**

*Yixuan Wang*1, ♦*Xuejing Zhao*1

1Lanzhou University

Risk prediction models quantify the probability of an individual experiencing specified events by modeling healthcare data, which is crucial for doctors to assess patients’ potential disease risks and provide personalized treatment. It has gathered significant

attention that the utilization of Longitudinal data during follow-up to dynamically update risk estimates in real-time in literatures. To incorporate with longitudinal covariates, existing dynamic prediction methods can be categorized into survival analysis and machine learning algorithms. Survival analysis methods are capable of predicting time-to-event outcomes and offer good model interpretability, but they often rely on strong assumptions, and involve high computational complexity. Meanwhile, even machine learning algorithms suitable for large-scale datasets exhibit powerful representation capabilities, they cannot model time-to-event outcomes and generally have lower predictive interpretability. This paper devotes to a dynamic prediction model that combines survival analysis and machine learning techniques, which integrates the deep learning network-DeepSurv into the landmarking framework to predict time-to-event outcomes with longitudinal covariates. Concretely, at each landmark time point, new information of the dataset is reconstructed, and the DeepSurv network is employed to capture the non-linear relationship between covariates and hazard rates, thereby give the prediction of survival functions. Furthermore, the importance of different features at each landmark time point is computed and ranked to evaluate the changing

impact of covariates on predictive outcomes over time. Performance on heart valve replacement surgery demonstrate that the proposed DeepSurv landmarking model outperforms state-of-art landmark methods, especially when the dependencies between covariates and the survival process are unknown or when multiple longitudinal covariates are present, which shows favorable predictive performance. Additionally, the model exhibits remarkable dynamic prediction capabilities in the real heart valve transplantation surgery.

**Session 24CHI49: New Developments in PCA Approaches**

**Modeling and Learning on High-Dimensional Matrix-Variate Sequences**

♦*Xu Zhang1, Catherine Chunling Liu2, Jianhua Guo3, K. C. Yuen4, A. H. Welsh5*

1South China Normal University，2The Hong Kong Polytechnic University，3Beijing Technology and Business University，4The University of Hong Kong， 5The Australian National University

We propose a new matrix factor model, named RaDFaM, which is strictly derived based on the general rank decomposition and assumes a structure of a high-dimensional vector factor model for each basis vector. RaDFaM contributes a novel class of low-rank latent structure that makes tradeoff between signal intensity and dimension reduction from the perspective of tensor subspace. Based on the intrinsic separable covariance structure of RaDFaM, for a collection of matrix-valued observations, we derive a new class of PCA variants for estimating loading matrices, and sequentially the latent factor matrices. The peak signal-to-noise ratio of RaDFaM is proved to be superior in the category of PCA-type estimations. We also establish the asymptotic theory including the consistency, convergence rates, and asymptotic distributions for components in the signal part. Numerically, we demonstrate the performance of RaDFaM in applications such as matrix reconstruction, supervised learning, and clustering, on uncorrelated and correlated data, respectively.

**Revisiting Asymptotic Theory for Principal Component Estimators of Approximate Factor Models**

♦*Peiyun Jiang*1, *Yoshimasa Uematsu*2, *Takashi Yamagata*3

1Tokyo Metropolitan University，2Hitotsubashi University，3University of York

It is well-known that the approximate factor models have the rotation indeterminacy. It has been considered that the principal component (PC) estimators estimate some rotations of the true factors and factor loadings, but the rotation matrix commonly used in the literature depends on the PC estimator itself. This raises a question: what does the PC estimator consistently estimate? This paper aims to explore the answer.

We first show that, assuming a quite general weak factor model with the $r$ signal eigenvalues diverging possibly at different rates, there always exists a unique rotation matrix composed only of the true factors and loadings, such that it rotates the true model to the identifiable model satisfying the standard $r^2$ restrictions. We call the rotated factors and loadings the pseudo-true parameters. We next establish the consistency and asymptotic normality of the PC estimator for this pseudo-true parameter. The results give an answer for the question: the PC estimator consistently estimates the pseudo-true parameter. We also investigate similar problems in the factor augmented regression. Finite sample experiments confirm the excellent approximation of the theoretical results.

**Principle Component Estimation for High-Dimensional Factor Models**

♦*Feifei Zhou*, *Zhigen Gao*, *Binghui Liu*

High-dimensional factor models have been widely applied in various fields. Principal component analysis (PCA) is a popular estimation method used to compute and provide consistent estimators for common factors and factor loadings in factor models. This paper contributes to the asymptotic properties of principal components estimates (PCE) in factor models when both the sample size (T) and the variable dimension (N) tend to infinity. Firstly, we propose bias-adjusted estimates of variance for both common factors and idiosyncratic errors. Secondly, we derive the convergence rate and limiting distributions for the common factors, factor loadings, and the variance of both common factors and idiosyncratic errors using a novel framework. Additionally, we conduct numerous numerical simulations to verify the convergence rate predicted by our theoretical results.

**High-Dimensional Vector Autoregression with Common Response and Predictor Factors**

♦*Di Wang1*

*1*Shanghai Jiao Tong University

The reduced-rank vector autoregressive (VAR) model can be interpreted as a supervised factor model, where two factor modelings are simultaneously applied to response and predictor spaces. This article introduces a new model, called vector autoregression with common response and predictor factors, to explore further the common structure between the response and predictors in the VAR framework. The new model can provide better physical interpretations and improve estimation efficiency. In conjunction with the tensor operation, the model can easily be extended to any finite-order VAR model. A regularization-based method is considered for the high-dimensional estimation with the gradient descent algorithm, and its computational and statistical convergence guarantees are established. For data with pervasive cross-sectional dependence, a transformation for responses is developed to alleviate the diverging eigenvalue effect. Moreover, we consider additional sparsity structure in factor loading for the case of ultra-high dimension. Simulation experiments confirm our theoretical findings and a macroeconomic application showcases the appealing properties of the proposed model in structural analysis and forecasting.

**Session 24CHI122: Statistical Methods of Survival Data**

**Estimation and Inference of Semiparametric Models for Directed Network Formation**

♦*Lianqiang Qu*1, *Lu Chen*, *Ting Yan* and *Yuguo Chen*

1Central China Normal University

We propose a semiparametric model for dyadic link formations in directed networks. The model contains a set of degree parameters that measure different effects of popularity or outgoingness across nodes, a regression parameter vector that reflects the homophily effect resulting from the nodal attributes or pairwise covariates associated with edges, and a set of latent random noises with unknown distributions. Our interest lies in inferring the unknown degree parameters and homophily parameters. The dimension of the degree parameters increases with the number of nodes. To address the issues of unknown noise distributions and the diverging dimensions of the parameter space, we develop a kernel-based least-squares approach to estimate the unknown parameters.

We prove consistency of all the resulting estimators of the degree parameters and homophily parameters. We establish high-dimensional central limit theorems for the proposed estimators and provide several applications of our general theory, including testing the existence of the degree heterogeneity, testing sparse signals and recovering the support. Finally, we demonstrate our theoretical results through simulation studies and present a real data analysis to illustrate their practical utility.

**Semiparametric Transformation Models of Survival-Out-of-Hospital**

♦*Xiaowei Sun*1, *Cheng Zeng*2 *and Liuquan Sun*2

1School of Mathematics and Statistics, Beijing Jiaotong University，2Institute of Applied Mathematics, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Recurrent event data with a terminal event commonly arise in biomedical studies, and the survival-out-of-hospital process is a useful alternative framework for the analysis of recurrent/terminal event data with non-negligible event duration. In this article, we propose a class of semiparametric transformation models for the survival-out-of-hospital process, and the proposed models offer great flexibility in formulating covariate effects on the probability of survival-out-of-hospital. Estimating equation approaches are developed for the model parameters, and the asymptotic properties of the resulting estimators are established. The finite sample performance of the proposed estimators is examined through simulation studies. An application to a Centers for Medicare and Medicaid Services study is provided.

**Nonparametric Statistical Inference via Metric Distribution Function in Metric Spaces**

*Xueqin Wang*1, *Jin Zhu*2, ♦*Wenliang Pan3*, *Junhao Zhu*2 *and Heping Zhang*4.

1University of Science and Technology of China， 2Sun Yat-Sen University， 3Chinese Academy of Sciences， 4Yale University

The distribution function is essential in statistical inference and connected with samples to form a directed closed loop by the correspondence theorem in measure theory and the Glivenko-Cantelli and Donsker properties. This connection creates a paradigm for statistical inference. However, existing distribution functions are defined in Euclidean spaces and are no longer convenient to use in rapidly evolving data objects of complex nature. It is imperative to develop the concept of the distribution function in a more general space to meet emerging needs. Note that the linearity allows us to use hypercubes to define the distribution function in a Euclidean space. Still, without the linearity in a metric space, we must work with the metric to investigate the probability measure. We introduce a class of metric distribution functions through the metric only.

We overcome this challenging step by proving the correspondence theorem and the Glivenko-Cantelli theorem for metric distribution functions in metric spaces, laying the foundation for conducting rational statistical inference for metric space-valued data. Then, we develop a homogeneity test and a mutual independence test for non-Euclidean random objects and present comprehensive empirical evidence to support the performance of our proposed methods.

**Semiparametric Representation Transfer Regression**

♦*Baihua He*1, *Huihang Liu*1 and *Jian Huang*2.

1USTC， 2Poly U

We propose a transfer learning method that utilizes data representations within the context of a semiparametric partially linear model. Our aim is to perform statistical inference on the parameter of primary interest in the target model while accounting for potential nonlinear effects of confounding variables. The influence of confounders can significantly affect the estimation of the main effect, and addressing the confounding effects of multiple covariates nonparametrically is challenging due to the curse of dimensionality. To circumvent this issue, we leverage knowledge from source domains, assuming that the sample size of the source data is substantially larger than that of the target data. This knowledge transfer is facilitated by the sharing of data representations, predicated on the idea that there exists a set of latent representations transferable from the source to the target domain. We address model heterogeneity between the source and target domains by incorporating domain-specific parameters in their respective models. We establish sufficient conditions for the identifiability of the models and demonstrate that the estimator for the primary parameter in the target model is both consistent and asymptotically normal. These results lay the theoretical groundwork for making statistical inferences about the main effects. Our simulation studies highlight the benefits of our method, and we further illustrate its practical applications using real-world data.

**Session 24CHI3: Advanced Statistical Approaches for Complex Data**

**Tyranny-of-the-Minority Regression Adjustment in Randomized Experiments**

*Xin Lu* *and* ♦*Hanzhong Liu*

Regression adjustment is widely used in the analysis of randomized experiments to improve the estimation efficiency of the treatment effect. This paper reexamines a weighted regression adjustment method termed tyranny-of-the-minority (ToM), wherein units in the minority group are given greater weights. We demonstrate that ToM regression adjustment is more robust than Lin (2013)'s regression adjustment with treatment-covariate interactions, even though these two regression adjustment methods are asymptotically equivalent in completely randomized experiments. Moreover, ToM regression adjustment can be easily extended to stratified randomized experiments and completely randomized survey experiments. We obtain the design-based properties of the ToM regression-adjusted average treatment effect estimator under such designs. In particular, we show that the ToM regression-adjusted estimator improves the asymptotic estimation efficiency compared to the unadjusted estimator, even when the regression model is misspecified, and is optimal in the class of linearly adjusted estimators. We also study the asymptotic properties of various heteroscedasticity-robust standard errors and provide recommendations for practitioners. Simulation studies and real data analysis demonstrate ToM regression adjustment's superiority over existing methods.

**Mining Multi-Brand Characteristics from Online Reviews for Competitive Analysis: A Brand Joint Model using Latent Dirichlet Allocation**

♦*Lu Xiaoling*

Online reviews reflect customers’ opinions of the products or services they have bought. Analysing online reviews provides a reliable way for e-commerce platforms to understand users’ needs and attitudes. To understand the strengths and weaknesses of several competitive brands, we propose a brand joint latent Dirichlet allocation model to analyse multi corpora simultaneously, particul s for multi-brre the general aspects of the online opinions of users concerning multi-brands and specific aspects of user opinions within individual brands. The results can assist the analysis of competitive brands and make meaningful suggestions for brand managers and marketers. We present two case studies to prove the efficiency of the proposed model.

**Penalized Sparse Covariance Regression with High Dimensional Covariates**

♦*Yuan Gao*1, *Zhiyuan Zhang*2, *Zhanrui Cai*3, *Xuening Zhu*2, *Tao Zou*4.

1Peking University，2Fudan University，3The University of Hong Kong, 4Australian National University

Covariance regression offers an effective way to model the large covariance matrix with the auxiliary similarity matrices. In this work, we propose a sparse covariance regression (SCR) approach to handle the potentially high-dimensional predictors (i.e., similarity matrices). Specifically, we use the penalization method to identify the informative predictors and estimate their associated coeﬀicients simultaneously. We first investigate the Lasso estimator and subsequently consider the folded concave penalized estimation methods (e.g., SCAD and MCP). However, the theoretical analysis of the existing penalization methods is primarily based on i.i.d. data, which is not directly applicable to our scenario. To address this diﬀiculty, we establish the non-asymptotic error bounds by exploiting the spectral properties of the covariance matrix and similarity matrices. Then, we derive the estimation error bound for the Lasso estimator and establish the desirable oracle property of the folded concave penalized estimator. Extensive simulation studies are conducted to corroborate our theoretical results. We also illustrate the usefulness of the proposed method by applying it to a Chinese stock market dataset.

**Model-Free joint Variable Selection via Importance Weighting**

♦*Yiwei Fan*1 *and**Junlong Zhao*2.

1School of Mathematics and Statistics, Beijing Institute of Technology， 2School of Statistics, Beijing Normal University

In numerous applications, modeling high-dimensional data poses the critical challenge of selecting significant variables. Various marginal screening methods have been proposed to tackle this challenge by evaluating the correlation between the response and predictor individually. However, such methods overlook the collective information of predictors. Another common strategy is forward selection, which involves selecting one significant variable in each iteration conditional on previously selected predictors. Although it represents an improvement over marginal methods, it still assesses correlations on a variable-by-variable basis, thus disregarding the interdependence among covariates outside the selected set.

To address these limitations and further enhance existing methods, we introduce a novel joint screening method. Drawing inspiration from techniques employed in deep neural networks, this method harnesses the collective information of predictors. Our approach aims to be applicable to both linear and nonlinear models, offering a promising avenue for improved variable selection in high-dimensional data analysis.

**Session 24CHI16: AI Frontiers in Clinical Data Analysis: Innovations and Applications**

**AI-Powered eCRT Package Translation**

♦*Zhanjian Dong* and *Yuzheng Ou*

Automating the translation of Electronic Case Report Tabulation (eCRT) packages into Chinese using ChatGPT. An eCRT package serves as a guide for regulatory reviewers, aiding their navigation through a study ‘s metadata. It is crucial for regulators to understand a clinical study data submitted for approval. In China, eCRT package are required to translate into Chinese which including annotated pdf CRF, Define.xml, SAS xpt, Reviewer’ Guide (sdrg, adrg). Currently, this task is leading to high costs and long delivery times. Traditional translation engines often struggle with translating eCRT algorithms that combine SAS code snippets, clinical terminologies, and complex logic, that can’t meet our acceleration requirements. We explored using Azure OpenAI GPT-4, with its ability to understand both text and codes and to learn and adapt to human instructions. The results are far exceeded our expectations.

**人工智能前沿技术在生命科学领域的应用**

♦*Ning Ma*

近十年来，以深度学习和大语言模型为代表的人工智能前沿技术快速发展，对各行各业产生了深远的影响。在生命科学领域，人工智能被广泛应用于基因组学、新药研发、精准医疗等细分板块，产生了像 Alphafold（蛋白质结构预测），Med-PaLM（医疗大模型）这样优秀的应用。国内外头部制药公司和医院也积极与 AI 公司合作，提升研发、临床试验、诊断和运营效率。

本主题环节综合介绍人工智能前沿技术在生命科学领域的应用，并探讨与之伴生的数据安全、隐私保护、技术鲁棒性与可信任问题。

**AI Assisted Clinical Data Validation**

♦*Bogong Zhu*

The clinical data for submission has a set of well-defined data standard, in which each variable has a clear definition of the purpose and the meaning. Furthermore, FDA or other regulartory agencies has defined a set of validation rules to make sure the submission data is above a standard bar. However, the rules are expressed in nature language and are updated regularly, making it time-consuming to follow.

Given the data standard and validation rules are publicly available, in the session I'll share the idea and some practice to leverage AI to learn the standard and understand the rules so that it can help the validation easier.

**Deep Networks as Denoising Algorithms: Sample-Efficient Learning of Diffusion Models in High-Dimensional Graphical Models**

♦*Song Mei*1, *Yuchen Wu*2

1UC Berkeley, 2University of Pennsylvania

We investigate the approximation efficiency of score functions by deep neural networks in diffusion-based generative modeling. While existing approximation theories utilize the smoothness of score functions, they suffer from the curse of dimensionality for intrinsically high-dimensional data. This limitation is pronounced in graphical models such as Markov random fields, common for image distributions, where the approximation efficiency of score functions remains unestablished.

To address this, we observe score functions can often be well-approximated in graphical models through variational inference denoising algorithms. Furthermore, these algorithms are amenable to efficient neural network representation. We demonstrate this in examples of graphical models, including Ising models, conditional Ising models, restricted Boltzmann machines, and sparse encoding models. Combined with off-the-shelf discretization error bounds for diffusion-based sampling, we provide an efficient sample complexity bound for diffusion-based generative modeling when the score function is learned by deep neural networks.

**Session 24CHI50: New Developments of Statistical Methods in Model Selection and Bio-Medical Studies**

**Estimation of Volume under the ROC Surface with Verification Biased Data**

*Shuangfei Shi1* and♦*Gengsheng Qin1*

1Georgia State University

In practice, the receiver operating characteristic (ROC) curve of a diagnostic test is widely used to show the performance of the test for discriminating two-class events. The area under the ROC curve (AUC) is proposed as an index for the assessment of the diagnostic accuracy of the test under consideration. Due to ethical and cost considerations associated with application of gold standard (GS) tests, only a subset of the patients initially tested have verified disease status. Statistical evaluation of the test performance based only on test results from subjects with verified disease status are typically biased. Various AUC estimation methods for tests with verification biased data have been developed over the last few decades. In this article, we develop new direct estimation methods for the volume under the ROC surface (VUS) by extending the AUC estimation methods for two-class diagnostic tests to three-class diagnostic tests in the presence of verification bias. The proposed methods will provide a comprehensive guide to deal with the verification bias in three-class diagnostic test accuracy studies and lead to a better choice of diagnostic tests.

**Variable Selection after Multiple Imputation on Epidemiological Data with Missing Values**

*Yang Li, Haoyu Yang, Haochen Yu, Hanwen Huang* and♦*Ye Shen*

Considering the inevitable correlation among different datasets within the same subject, we propose a framework of variable selection on multiply imputed data with penalized weighted least squares (PWLS–MI). The methodological development is motivated by an epidemiological study of A/H7N9 patients from Zhejiang province in China, where nearly half of the variables are not fully observed. Multiple imputation is commonly adopted as a missing data processing method. However, it generates correlations among imputed values within the same subject across datasets. Recent work on variable selection for multiply imputed data does not fully address such similarities. We propose PWLS–MI to incorporate the correlation when performing the variable selection. PWLS–MI can be considered as a framework for variable selection on multiply imputed data since it allows various penalties. We use adaptive LASSO as an illustrating example. Extensive simulation studies are conducted to compare PWLS–MI with recently developed methods and the results suggest that the proposed approach outperforms in terms of both selection accuracy and deletion accuracy. PWLS–MI is shown to select variables with clinical relevance when applied to the A/H7N9 database.

**Sparse Maximum Likelihood Estimation for Regression Models**

♦*Min Tsao1*

1University of Victoria, Canada

For model selection under the maximum likelihood framework, we study the likelihood ratio confidence region for the regression parameter vector of a full regression model. We show that when its confidence level increases with the sample size at a certain speed, with probability tending to one, the confidence region consists of vectors representing models containing all active variables, including the true parameter vector of the full model. Using this result, we devise a consistent model selection criterion which has a sparse maximum likelihood interpretation and certain advantages over popular information criteria. We also provide a large-sample characterization of models of maximum likelihood at different model sizes which shows that, for selection consistency, it suffices to consider only this small set of models.

**False Discovery Rate Control for Variable Selection in Spatial Linear Models**

*Yang Wang1, Yingke Yang1* and♦*Zhouping Li1*

1Lanzhou University

Selecting important variables is a crucial task in statistical modeling. In this talk, we discuss the problem of selecting important covariates in spatial linear models with conditional autoregressive (CAR) and the simultaneous autoregressive (SAR) errors. Specifically, in order to control the false discovery rate (FDR) in the variable selection procedure of the spatial linear models, we propose a novel knockoff filter under this setup. We prove that the proposed method is able to control the FDR under finite sample sizes. We also illustrate the finite sample performance of the proposed selection approach via simulation studies and real data analysis.

**Session 24CHI75: Recent Advances in Statistical Methods for Complex Data Analysis**

**Dynamic Supervised Principal Component Analysis for Classification**

*Wenbo Ouyang1, Ruiyang Wu2,* ♦*Ning Hao* and *Hao1 Helen Zhang1*

1The University of Arizona，2New York University

In this talk we introduce a novel framework for dynamic classification in high dimensional spaces, addressing the evolving nature of class distributions over time or other index variables. Traditional discriminant analysis techniques are adapted to learn dynamic decision rules with respect to the index variable. We propose and study a new supervised dimension reduction method employing kernel smoothing to identify the optimal subspace and provide a comprehensive examination of this approach for both linear discriminant analysis and quadratic discriminant analysis. We illustrate the effectiveness of the proposed methods through numerical simulations and real data examples. The results show considerable improvements in classification accuracy and computational efficiency.

**Clustered Coefficient Regression Models for Poisson Process with an Application to Seasonal Warranty Claim Data**

*Xin Wang1, Xin Zhang2* and♦*Zhengyuan Zhu3*

1San Diego State Universitity， 2Meta， 3Iowa State University

Motivated by a product warranty claims dataset, we propose clustered coefficient regression models in a nonhomogeneous Poisson process for recurrent event data. The proposed method, referred to as CLUPP, can estimate the group structure and parameters simultaneously. In our proposed method, a penalized regression approach is used to identify the group structure. Numerical studies show that the proposed approach can identify the group structure well, and outperforms traditional methods such as hierarchical clustering and K-means. We also establish theoretical properties, which show that the proposed estimators can converge to true parameters in high probability. In the end, we apply our proposed methods to the product warranty claims dataset, which achieve better prediction than the state-of-the-art methods.

**Bayesian Nonparametrics Models for Grouped Data: Plaid Atoms Model**

♦*Yuan Ji1* and *Dehua Bi1*

1The University of Chicago

We consider a new class of Bayesian nonparametrics models called Plaid Atoms Model (PAM) for grouped data. The goal is to reveal complex structure of heterogeneity across experimental units like dependent clusters. Under principled and model-based inference, PAM is able to report shared and unique subpopulations as dependent clusters across grouped data sets. Through simulation and case studies, we demonstrate the performance of PAM in comparison with other related models.

**Linear Algorithms for Robust and Scalable Nonparametric Multiclass Probability Estimation**

*Liyun Zeng1* and♦*Hao Zhang1*

1University of Arizona

Multiclass probability estimation is the problem of estimating conditional probabilities of a data point belonging to a class given its covariate information. It has broad applications in statistical analysis and data science. Recently a class of weighted Support Vector Machines (wSVMs) has been developed to estimate class probabilities through ensemble learning for K-class problems, where K is the number of classes. The estimators are robust and achieve high accuracy for probability estimation, but their learning is implemented through pairwise coupling, which demands polynomial time in K. In this paper, we propose two new learning schemes, the baseline learning and the One-vs-All (OVA) learning, to further improve wSVMs in terms of computational efficiency and estimation accuracy. In particular, the baseline learning has optimal computational complexity in the sense that it is linear in K. Though not being most efficient in computation, the OVA offers the best estimation accuracy among all the procedures under comparison. The resulting estimators are distribution-free and shown to be consistent. We further conduct extensive numerical experiments to demonstrate finite sample performance.

**Session 24CHI94: Recent Developments of Statistical and Computational Methodologies for Genomic Data**

**An Adaptive Fisher’s Method for Genome-Wide Multi-Trait Association Study with Summary Statistics**

♦*Shili Lin1*

1Ohio State University

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with traits. Commonly used single-trait analysis methods are conservative, while multi-trait methods improve statistical power by integrating multiple association evidence. In contrast to individual-level data, GWAS summary statistics are usually publicly available, and thus methods using only summary statistics have greater usage. Although many methods have been developed for joint analysis of multiple traits using summary statistics, there exit issues such as inconsistent performance, computational inefficiency, and numerical problems when considering a large number of traits. In this work, we develop an adaptive Fisher method MTAFS, a summary statistics-based multi-trait analysis method that is powerful, robust, and computationally efficient. We demonstrate the utility of MTAFS by applying it to two sets of brain imaging derived phenotypes (IDPs) from the UK Biobank. We provide evidence to show that the underlying genes of the single nucleotide polymorphisms identified by MTAFS are found to exhibit higher expression and are significantly enriched in brain-related tissues.

**Leverage Genome-Wide SNPs and Biobank-Scale Data for Efficient and Improved Estimation of Genetic Admixture Proportions**

♦*Baolin Wu1*

1University of California - Irvine

Genetic admixture estimation has been widely studied and proven very useful in ancestry inference and genome-wide association studies.

Existing methods and tools, such as ADMIXTUER and OpenADMIXTURE, have largely relied on modeling a subset of independent SNPs, partly for computational convenience. In this talk, we present an innovative and scalable statistical modeling framework to improve the accuracy of genetic admixture proportion estimation. We will briefly discuss (1) a novel semi-supervised modeling approach that can leverage outside biobank summary data just using allele frequency; (2) a new admixture model that can efficiently model genome-wide genetic markers for much improved inference without need of pruning, (3) a new penalized model that can seamlessly rank/select important ancestry informative markers. These newly proposed approaches were applied to public GWAS data to demonstrate their competitive performance.

We have implemented the proposed methods in an open source R package.

**An Explicit Method for SNP-Heritability Estimation using GWAS Summary Statistics without Heritability Modeling**

♦*Kai Wang1.*

1University of Iowa

There are quite a few methods for estimating SNP-heritability with summary statistics. Most of them are based on mixed effects models with additional modeling on heritability. Here I propose an explicit method for estimating SNP-heritability from GWAS summary statistics. This method is based on the fixed-effect linear model and does not need heritability modeling. Both point and interval estimates of the SNP-heritability are obtained from the non-central $F$ distribution. This method is applied to the GIANT consortium and UK BioBank Meta-analysis summary statistics on height and BMI (average sample size 700,000). The point and interval estimates are also applicable to individual-level data. The proposed framework naturally allows for power and sample size calculations regarding SNP heritability.

**Robustness and Resilience of Computational Deconvolution Methods for bulk RNA Sequencing Data**

*Su Xu1, Duan Chen1, Xue Wang2* and ♦*Shaoyu Li1*

1University of North Carolina at Charlotte， 2Mayo Clinic

In this study, we simulate pseudo bulk RNA sequencing data to benchmark the robustness and resilience of reference-based and reference-free bulk tissue RNA deconvolution methods. In silico pseudo bulk tissue RNA sequencing data to test the methods’ robustness under violated assumptions. Additionally, resilience is evaluated by altering single cell profiles distributions in generating the pseudo bulk tissue RNAseq data. Four distinct single cell RNA (scRNA) sequencing datasets from various tissues types are employed. Estimations are compared to the ground truth using Pearson correlation coefficient, root mean squared deviation and mean absolute deviation. Results indicate that reference-based methods display greater robustness when reliable reference scRNA sequencing data are available., while, reference-free methods exhibit strengths in the absence of appropriate scRNA reference data.

**Session 24CHI105: Spatial Data Modeling and Applications in Biomedical Data Analysis**

**Differential Expression Analysis of Spatial Transcriptomic Data**

*Fei Qin1, Xizhi Luo2, Bo Cai1, Feifei Xiao3* and ♦*Guoshuai Cai3*

1University of South Carolina，2AbbVie Inc.， 3University of Florida

The emergence of spatial transcriptomic technologies has opened new avenues to investigate gene activities while preserving the spatial context of tissues. However, current methods available for analyzing spatial transcriptomic data are still in their infancy, and none of the existing methods is capable of identifying SV genes between groups. To overcome this challenge, we developed SPADE for spatial pattern and differential expression analysis to identify SV genes in spatial transcriptomic data. Through extensive simulations and real data analyses, we have demonstrated the superior performance of SPADE compared to existing methods in detecting spatially variable (SV) genes within and between groups.

**ENGEP: Advancing Spatial Transcriptomics with Accurate Unmeasured Gene Expression Prediction**

*Shi-Tong Yang1* and♦*Xiao-Fei Zhang1*

1Central China Normal University

Imaging-based spatial transcriptomics techniques provide valuable spatial and gene expression information at single-cell resolution. However, their current capability is restricted to profiling a limited number of genes per sample, resulting in most of the transcriptome remaining unmeasured. To overcome this challenge, we develop ENGEP, an ensemble learning-based tool that predicts unmeasured gene expression in spatial transcriptomics data by using multiple single-cell RNA sequencing datasets as references. ENGEP outperforms current state-of-the-art tools and brings biological insight by accurately predicting unmeasured genes. ENGEP has exceptional efficiency in terms of runtime and memory usage, making it scalable for analyzing large datasets.

**Bayesian Semiparametric Modeling for Spatial Partly Interval-Censored Data**

♦*Bo Cai1* and *Chun Pan2*.

1University of South Carolina， 2Hunter College

Partly interval-censored data often occur in biomedical studies of diseases where periodic examinations for symptoms of interest are necessary. Subjects' geographic information (e.g. county, latitude/longitude) is sometimes also available and can be useful in predicting survival. We propose a Bayesian semiparametric method for analyzing partly interval-censored data with areal spatial information under the proportional hazards model. A simulation study is conducted to compare the performance of the proposed method with the main method currently available in the literature. The method is illustrated through a geographyically referenced leukemia survival data. The proposed method will be especially useful for analyzing progression-free survival in multi-regional cancer clinical trials.

**AI Powered Spatial Transcriptomics for Knowledge Discovery**

♦*Qianqian Song1*

1University of Florida

Spatial omics aims to profile different molecular layers of omics profiles within a tissue's spatial context, providing insights into cell behavior, interactions, and function. The application of AI techniques to this field offers unique opportunities for analyzing multi-modality data such as histological image, single cell omics, and spatial cellular relations, thus deciphering spatial patterns and making biologically significant discoveries. The session will address how AI can help us handle the complexity of spatial omics data, uncover hidden biological insights, and facilitate our understanding of tissue microenvironment. This presentation aims to delve into the intersection of Artificial Intelligence (AI) and Spatial Omics—a cutting-edge area in bioinformatics. Spatial Omics is rapidly gaining attention for its potential to revolutionize our understanding of cellular activities within the context of tissue architecture. This session will showcase the recent advances, challenges, and the immense potential of AI-driven spatial omics research.

**Session 24CHI112: Statistical Considerations of Time-to-Event Analysis in Modern Oncology Clinical Trials**

**Case Sharing Reflections for Time-to-Event Sensitivity Analysis from a Confirmatory Oncology Clinical Trial with Open-label Design**

♦*Yuqi Wang1, Gang Cheng1*

1Jiangsu Hengrui Medicine CO.,LTD

The CodeBreaK 200 trial, a randomized, open-label, parallel-controlled phase 3 cilical trial, aimed to test Sotorasib, the first orally selective covalent inhibitor of KRAS G12C, as a treatment for people with advanced NSCLC. The trial showed that Sotorasib had a limited improvement to PFS (5.6 months v.s 4.5 months) and showed no statistically significant difference in OS compared to Docetaxel. Based on the results, the Oncology Drugs Advisory Committee (ODAC) concluded that due to the open-label design, asymmetric early withdrawal between groups, potential misuse of the Confirmation of Progression (COP) procedure, and discordance in assessment of progressive disease between investigators and BICR would have led to systematic bias in the trial, resulting in the trial results failing to demonstrate the reliability of PFS. We further evaluated the influence of asymmetric early dropouts on PFS via simulation, especially when the patients with better baseline status in the control group were more likely to switch therapy early. When this asymmetric dropouts happened, the estimated PFS for the control group tended to be shorter than expected, which would lead to the inflation of type I error. The earlier the dropouts and the higher the dropout rate, the greater the impact on PFS it would casue. The undifferencentied groups, which would have shown no statistically different results, will be more likely to demostrate the improvement in PFS with such asymmetric early dropouts. Similar to CodeBreak 200, when 15% of “good” patients in the control group dropped within 1.25 months after randomization, there would be a probability of 27% to show a significant increase of 1 month in PFS even both groups had the same PFS of 6.5 months without dropouts. If this happens, the results should be interpreted with caution and the potential systemic bias should be thoroughly evaluated.

**Performance of Right-Censored vs. Interval-Censored Methods for Evaluating Time to Progression and Progression-Free Survival.**

♦*Zhihao Cheng*1, *Tao Zhang*1, *Zhao Wang*1，*Yang Yan1*

1Qilu Pharmaceutical Co

Background: Tumor progression, occurring between two consecutive imaging evaluations, represents interval-censored data. Traditional pharmaceutical analyses utilize right-censored methods (e.g., Kaplan-Meier curve, log-rank test, Cox model) to analyze Time to Progression (TTP) and Progression-Free Survival (PFS), which may lead to bias. Objective: Our research aims to uncover the biases in estimating median TTP (mTTP), median PFS (mPFS), hazard ratio (HR), and statistical power using right-censored analysis and demonstrate the efficacy of interval-censored analysis in mitigating these biases. Methods: Employing Monte Carlo simulations, we evaluated the biases in TTP and PFS analyses using five methods across varied scenarios. These methods include right-censoring analyses using (1) the date of imaging progression, (2) the date of the last imaging before progression, (3) the midpoint between the date of imaging progression and the last imaging date, and (4) the exact progression time (idealistic but not feasible in reality) as the event time, respectively; and (5) Interval-censored analysis. Scenarios varied in imaging intervals, consistency of imaging intervals between experimental and control groups, and actual median progression time and death time. Results: (1) For the TTP endpoint, Method 1 overestimated mTTP, though having minimal and negligible biases on HR estimates and statistical power, when the imaging intervals were consistent between comparison groups. Inconsistent imaging intervals, however, led to apparent biases in HR and statistical power. Method 2 displayed a bias opposite to Method 1, while Method 3’s bias direction was uncertain. Methods 4 and 5 provided accurate estimates for mTTP, HR, and statistical power across scenarios. (2) For the PFS endpoint, Method 1 similarly biased mPFS estimation. Notably, in scenarios where the HR for tumor progression was lower than for death and the median times of the two events were close (e.g., post-line treatments), consistent imaging intervals did not prevent biases in HR and statistical power under Method 1, whereas Method 5 (interval-censored analysis) remained reliable. Conclusion: Traditional right-censored analyses, using the date of imaging progression as the progression time, overestimate mTTP and mPFS, leading to skewed HR estimates and statistical power in some realistic scenarios. Interval-censored analysis successfully rectifies these biases.

**Event-Free Survival As a Surrogate Endpoint for Overall Survival in Previously Untreated Acute Myeloid Leukemia**

♦*Jun Yin*1

1Moffitt Cancer Center

Overall survival (OS) remains the definitive primary efficacy endpoint to evaluate previously untreated acute myeloid leukemia (AML) therapies, but it requires prolonged follow-up. An earlier endpoint assessed post-treatment would expedite clinical trial conduct and accelerate patient access to effective new therapies. Our objective was to formally evaluate event-free survival (EFS) as a surrogate endpoint for OS in untreated AML. Individual patient data were analyzed from 2,475 patients (pts) from 4 multicenter, randomized controlled phase III trials of active treatment in previously untreated AML using anthracycline and cytarabine induction chemotherapy as the concurrent control. Individual patient-level surrogacy examines the association between the individual patients' EFS and OS time after adjusting for treatment effect, and trial-level surrogacy measures how precisely the treatment effect on OS can be predicted on the basis of observed treatment effect on EFS. Despite the lack of patient-level correlation, a strong correlation between hazard ratios for treatment effects was observed between EFS and OS on the trial level. Hence, it remains debatable whether EFS represents a clinical benefit in itself for patient with untreated AML considering the strong correlation in treatment effects. Further validation is needed due to the small number of trials included and the heterogeneity across trials.

**Covariate Adjustment or Stratified Analyses for Time-to-Event Endpoints**

♦*Xiuyu Cong*

Covariate adjustment in analyzing randomized controlled clinical trials has been of ever-growing interest. Recently in 2023 the US FDA issued draft guidance that emphasizes the efficiency and precision gained by adjusting for prognostic covariates. This presentation will provide a systemic review of covariate adjustment in statistical testing and estimation with a focus on covariate-adjusted log-rank test and proportional hazard model, in contrast to their stratified counterparts that is currently more commonly used. If appropriate, method used for covariate-adaptive randomization will also be discussed. Case studies will be provided to facilitate discussion.

**Session 24CHI11: Advances in Kernel Methods and Nonparametric Inference for Complex Data**

**Optimal Rate of Kernel Regression in Large Dimensions**

♦*Weihao Lu1, Haobo Zhang1, Yicheng Li1, Manyun Xu1，* *Qian Lin1*

1Tsinghua University

We perform a study on kernel regression for large-dimensional data (where the sample size $n$ is polynomially depending on the dimension $d$ of the samples, i.e., $n\asymp d^{\gamma}$ for some $\gamma >0$ ).

 We first build a general tool to characterize the upper bound and the minimax lower bound of kernel regression for large dimensional data through the Mendelson complexity $\varepsilon\_{n}^{2}$ and the metric entropy $\bar{\varepsilon}\_{n}^{2}$ respectively. When the target function falls into the RKHS associated with a (general) inner product model defined on $\bbS^{d}$, we utilize the new tool to show that the minimax rate of the excess risk of kernel regression is $n^{-1/2}$ when $n\asymp d^{\gamma}$ for $\gamma =2, 4, 6, 8, \cdots$. We then further determine the optimal rate of the excess risk of kernel regression for all the $\gamma>0$ and find that the curve of optimal rate varying along $\gamma$ exhibits several new phenomena including the \textit{multiple descent behavior} and the \textit{periodic plateau behavior}.As an application, For the neural tangent kernel (NTK), we also provide a similar explicit description of the curve of optimal rate. As a direct corollary, we know these claims hold for wide neural networks as well.

**AMMD-Based New Test for Equality of Several Distributions in Separable Metric Spaces**

♦*Jin-Ting Zhang1, Jia Guo2 ， Bu Zhou3*

1Department of Statistics and Data Science, National University of Singapore，2School of Management, Zhejiang University of Technology, China，3School of Statistics and Mathematics, Zhejiang Gongshang University, China

A new test for equal distributions of several high-dimensional samples in separable metric spaces, with its test statistic constructed based on maximum mean discrepancy, is proposed and studied. The asymptotic null distribution of the test statistic is established under some mild conditions. The new test is implemented via a three-cumulant matched chi-square approximation with the associated approximation parameters consistently estimated from the data. A new data-adaptive Gaussian kernel width selection method is also suggested. Good performance of the new test is illustrated by intensive simulation studies and a real data example of Gini index curves.

**The Optimality of Kernel Classifiers in Sobolev Space**

*Jianfa Lai1, Zhifan Li1,* ♦*Dongming Huang2* and *Qian Lin1*

1Tsinghua University，2National University of Singapore

Kernel methods are widely used in machine learning, especially for classification problems. However, the theoretical analysis of kernel classification is still limited. This paper investigates the statistical performances of kernel classifiers. With some mild assumptions on the conditional probability η(x)=ℙ(Y=1 | X=x), we derive an upper bound on the classification excess risk of a kernel classifier using recent advances in the theory of kernel regression. We also obtain a minimax lower bound for Sobolev spaces, which shows the optimality of the proposed classifier. Our theoretical results can be extended to the generalization error of overparameterized neural network classifiers. To make our theoretical results more applicable in realistic settings, we also propose a simple method to estimate the interpolation smoothness of 2η(x)−1 and apply the method to real datasets.

**Scalable Statistical Inference in Non-parametric Least Squares**

♦*Meimei Liu*1, *Zuofeng Shang*2 *and Yun Yang*3.

1Virginia Tech, 2NJIT，3UIUC.

Stochastic approximation (SA) is a powerful and scalable computational method for iteratively estimating the solution of optimization problems in the presence of randomness, particularly well-suited for large-scale and streaming data settings. In this work, we propose a theoretical framework for stochastic approximation (SA) applied to non-parametric least squares in reproducing kernel Hilbert spaces (RKHS), enabling online statistical inference in non-parametric regression models. We achieve this by constructing asymptotically valid pointwise (and simultaneous) confidence intervals (bands) for local (and global) inference of the nonlinear regression function, via employing an online multiplier bootstrap approach to a functional stochastic gradient descent (SGD) algorithm in the RKHS. Our main theoretical contributions consist of a unified framework for characterizing the non-asymptotic behavior of the functional SGD estimator and demonstrating the consistency of the multiplier bootstrap method. Our method specifically reveals an interesting relationship between the tuning of step sizes in SGD for estimation and the accuracy of uncertainty quantification.

**Session 24CHI26: Computational and Methodological Statistics**

**Causal Mediation Analysis for Multilevel and Functional Data**

*Yi Zhao1,* ♦*Xi Luo2, Michael Sobel3, Martin Lindquist4 and Brian Caffo4.*

1Indiana University， 2Univ of Texas Health Science Center at Houston， 3Columbia University， 4Johns Hopkins University

Causal mediation analysis typically involves conditions that may not be applicable in neuroimaging studies. We introduce a multilevel causal mediation framework to overcome this limitation and more accurately quantify information flow in brain pathways. This framework is designed to tackle several challenges: unmeasured mediator-outcome confounding, multilevel time series analysis, and the estimation of functional causal effects. Our approach is grounded in multilevel structural equation modeling, complemented by relaxed likelihood estimation methods. Interestingly, certain causal estimates, typically unobtainable in simpler data structures, become identifiable in our more complex data setting. We provide proof of the asymptotic properties of our estimators and illustrate the numerical properties through empirical analysis. Additionally, we utilize real fMRI data to demonstrate the practical effectiveness of our proposed framework.

**Distributionally Robust Markov Decision Process with Uncertain Transition Probabilities**

♦*Yue Shi*1 and *Yisha Xiang*2

1Wuhan University, 2University of Houston

To mitigate the effects of parameter uncertainty due to statistical estimation errors, many studies use data-driven methods to construct an ambiguity set that contains the true parameters with high confidence and formulate the problem of interest as a robust Markov decision process that helps decision makers hedge against worst-case parameters. An inherent drawback associated with this approach is the difficulty in incorporating priori probabilistic information of unknown parameters, which can potentially be leveraged to find a robust yet not too conservative policy. In this study, we model the uncertainty in transition probabilities using a multiple-priors model and propose a distributionally robust Markov decision process (DRMDP) with a multiple-priors-based uncertainty set. We investigate the asymptotic convergence of the posterior set and further examine the asymptotic convergence of the optimal value and optimal policy of the proposed DRMDP model. Moreover, we develop an efficient approximation method to solve the proposed DRMDP model and analytically examine the approximation performance. Numerical experiments on a machine replacement problem show the benefit of the proposed decision-making framework in providing good out-of-sample performance.

**Bayesian Mixed-Effect Higher-Order Hidden Markov Models with Applications to Predictive Healthcare using Electronic Health Records**

♦*Ying Liao*1, *Yisha Xiang*2, *Zhigen Zhao*3, *Di Ai*4

1Wuhan University, 2University of Houston , 3Temple University, 4The University of Texas Health Science Center at Houston

The disease progression dynamics in electronic health records often reflect patients' health condition evolution, holding the promise of enabling the development of clinical predictive models. These dynamics, however, generally exist significant variability among patients due to some critical factors (e.g., gender and age) and patient-level heterogeneity. Moreover, future health state may not only depend on the current state but also more distant history states due to the complicated disease progression. To capture this complex transition behavior and address mixed effects in clinical prediction problems, we propose a novel and flexible Bayesian mixed-effect higher-order hidden Markov model (MHOHMM), and develop a classifier based on MHOHMMs. A range of MHOHMMs are designed to capture different data structures and the optimal one is identified by using the k-fold cross-validation. An effective two-stage Markov chain Monte Carlo (MCMC) sampling algorithm is designed for model inference. A simulation study is conducted to evaluate the performance of the proposed sampling algorithm and the MHOHMM-based classification method. The practical utility of the proposed framework is demonstrated by a case study on the acute hypotensive episode prediction for intensive care unit patients. Our results show that the MHOHMM-based framework provides good prediction performance.

**On Partial Envelope Approach for Modeling Spatial-Temporally Dependent Data**

♦*Wenbo Wu*1, *Reisa Widjaja*1, *Keying Ye*1

1The University of Texas at San Antonio

In the new era of big data, modeling multivariate spatial-temporally dependent data is a challenging task due to the dimensionality of the features and complex spatial-temporal associations among the observations across different locations and time points. To improve the estimation efficiency, we propose a spatial-temporal partial envelope model which is parsimonious and effective in modeling high-dimensional spatial-temporal data. The partial envelope model was proposed under a linear coregionalization model framework which allows heterogenous spatial-temporal covariance structure for different components of the response vector. The maximum likelihood estimator for the proposed model can be obtained through a Grassmann manifold optimization. We obtained a complete asymptotic result for the estimator and conduct thorough empirical simulations to demonstrate the soundness and effectiveness of the proposed method. We also apply the proposed model to analyze the crowdsourcing weather data collected from personal weather stations in the city of San Antonio, TX of the United States.

**Session 24CHI37: High-Dimensional Data Analysis and Variable Selection**

**CoxKnockoff: Controlled Feature Selection for the Cox Model Using Knockoffs**

♦*Daoji Li1, Jinzhao Yu2 and Hui Zhao2*

1California State University, Fullerton， 2Zhongnan University of Economics and Law

Although there is a huge literature on feature selection for the Cox model, none of the existing approaches can control the false discovery rate (FDR) unless the sample size tends to infinity. In addition, there is no formal power analysis of the knockoffs framework for survival data in the literature. To address those issues, in this paper, we propose a novel controlled feature selection approach using knockoffs for the Cox model. We establish that the proposed method enjoys the FDR control in finite samples regardless of the number of covariates. Moreover, under mild regularity conditions, we also show that the power of our method is asymptotically one as sample size tends to infinity. To the best of our knowledge, this is the first formal theoretical result on the power for the knockoffs procedure in the survival setting. Simulation studies confirm that our method has appealing finite-sample performance with desired FDR control and high power. We further demonstrate the performance of our method through a real data example. This is a joint work with Jinzhao Yu and Hui Zhao.

**Self-Normalized Cram\'er Type Moderate Deviation Theorem for Gaussian Approximation**

♦*Jingkun Qiu1, Song Xi Chen2 and Qi-Man Shao2.*

1Guanghua School of Management, Peking University， 2Guanghua School of Management and Center for Statistical Science, Peking University， 3Department of Statistics and Data Science, SICM, National Center for Applied Mathematics Shenzhen, Southern University of Science and Technolog Berry--Esseen type bounds for Gaussian approximation of standardized sums have been extensively studied under exponential type moment conditions.

In this paper, a Cram\'er type moderate deviation theorem is established for self-normalized Gaussian approximation under finite moment conditions.More specifically, let $X\_{1},X\_{2},\dots,X\_{n}$ be i.i.d. $\mathbb{R}^{p}$-valued random vectors with zero means.Let $S\_{n,j}=\sum\_{i=1}^{n}X\_{ij}$ and $V\_{n,j}^{2}=\sum\_{i=1}^{n}X\_{ij}^{2}$.We show that if the correlation matrix of $X\_{1}$ is $I\_{p}$ and the third moment of $X\_{1}$ is finite, then\[\frac{\mathbb{P}(\max\_{1\leq j\leq p}S\_{n,j}/V\_{n,j}>x)}{\mathbb{P}(\max\_{1\leq j\leq p}Z\_{j}>x)}\to1\]uniformly for $0\leq x\leq o(n^{1/6})$ and for all $p\geq1$, where $Z\_{1},\dots,Z\_{p}$ are independent standard normal random variables.Similar result is also established for large $x$ when $X\_{1}$ has a general correlation matrix.The proof is based on a new Cram\'er type moderate deviation theorem for the minimum of several self-normalized sums.

**Enhancing Efficiency and Robustness in High-Dimensional Linear Regression with Additional Unlabeled Data**

♦*Yuqian Zhang1，Kai Chen1.*

1Renmin University of China

In semi-supervised learning, the prevailing understanding suggests that observing additional unlabeled samples improves estimation accuracy for linear parameters only in the case of model misspecification. This paper challenges this notion, demonstrating its inaccuracy in high dimensions. Initially focusing on a dense scenario, we introduce robust semi-supervised estimators for the regression coefficient without relying on sparse structures in the population slope. Even when the true underlying model is linear, we show that leveraging information from large-scale unlabeled data improves both estimation accuracy and inference robustness. Moreover, we propose semi-supervised methods with further enhanced efficiency in scenarios with a sparse linear slope. Diverging from the standard semi-supervised literature, we also allow for covariate shift. The performance of the proposed methods is illustrated through extensive numerical studies, including simulations and a real-data application to the AIDS Clinical Trials Group Protocol 175 (ACTG175).

**Semi-Parametric Tensor Factor Analysis by Iteratively Projected Singular Value Decomposition**

♦*Elynn Chen*, *Dong Xia*, *Chencheng Cai* ， *Jianqing Fan*.

This paper introduces a general framework of Semi-parametric TEnsor Factor Analysis (STEFA) that focuses on the methodology and theory of low-rank tensor decomposition with auxiliary covariates. Semi-parametric TEnsor Factor Analysis models extend tensor factor models by incorporating auxiliary covariates in the loading matrices. We propose an algorithm of iteratively projected singular value decomposition (IP-SVD) for the semi-parametric estimation. It iteratively projects tensor data onto the linear space spanned by the basis functions of covariates and applies singular value decomposition on matricized tensors over each mode. We establish the convergence rates of the loading matrices and the core tensor factor. The theoretical results only require a sub-exponential noise distribution, which is weaker than the assumption of sub-Gaussian tail of noise in the literature. Compared with the Tucker decomposition, IP-SVD yields more accurate estimators with a faster convergence rate. Besides estimation, we propose several prediction methods with new covariates based on the STEFA model. On both synthetic and real tensor data, we demonstrate the efficacy of the STEFA model and the IP-SVD algorithm on both the estimation and prediction tasks.

**Session 24CHI83: Recent Development in Change-**

**Point Detection Problem**

**Graph-Based Multiple Change-point Detection**

*Yuxuan Zhang，*♦*Hao Chen1*

1University of California, Davis

We propose a new multiple change-point detection framework for multivariate and non-Euclidean data. First, we combine graph-based statistics with wild binary segmentation or seeded binary segmentation to search for a pool of candidate change-points. We then prune the candidate change-points through a novel goodness-of-fit statistic. Numerical studies show that this new framework outperforms existing methods under a wide range of settings. The resulting change-points can further be arranged hierarchically based on the goodness-of-fit statistic. The new framework is illustrated on a Neuropixels recording of an awake mouse.

**Spectral Norm of Exponentially Weighted Moving Sample Covariance Matrix and Its Application to Sequential Sparse Signal Detection**

♦*YanHong Wu1, wei biao wu2* and *Dong-Yun Kim3*

1California State University Stanislaus， 2University of Chicago， 3NIH

This study focuses on the exponentially weighted moving sample covariance matrix, investigating its behavior in both null and alternative hypotheses. Under the null hypothesis, assuming normal observations, we establish exponential probability bounds for the largest

eigenvalue. Similarly, under the alternative hypothesis with a single spike, we derive corresponding bounds. Notably, as the smoothing weight parameter tends to zero, our bounds ensure almost sure convergence. We extend our findings to sub-gaussian vectors, providing

exponential bounds for the largest eigenvalue in terms of both the weight parameter and the norm of sub-gaussian random variables. Furthermore, we address the sparse signal case by adapting our techniques as the dimension of observation approaches infinity. In the context of almost sure convergence, we present results for scenarios where the product of dimension and weight parameter converges to a constant. This comprehensive exploration enhances our understanding of the behavior of the exponentially weighted moving sample covariance matrix in various statistical settings. These results are then used to compare several sequential monitoring charts for covariance matrix change by using the spectrum of EWMV

**A Change-point Method for Phase I Analysis of High-dimensional Processes with Sparse Mean Shifts**

♦*Lianjie Shu*

Although Phase I analysis of multivariate processes has been extensively discussed, the discussion on techniques for Phase I monitoring of high-dimensional processes is still limited. In high-dimensional applications, it is common to observe that a large number of components but only limited of them change at the same time. The shifted components are often sparse and unknown a priori in practice. Motivated by this, this paper studies Phase I monitoring of high-dimensional process mean vectors under an unknown sparsity level of shifts. The basic idea of the proposed control chart is to first employ the false discovery rate (FDR) procedure to estimate the sparsity level of mean shifts, and then to monitor the mean changes based on the maximum of the directional likelihood ratio (DLR) statistics over all the possible shift directions. The comparison results based on extensive simulations favor the proposed control chart. A real example is presented to illustrate the implementation of the new control chart.

**Constructing Generalized Optimal Control Charts for Detecting Distribution Changes in Dependent Observations Sequence**

♦*Dong Han*1, *Fugee Tsung*2

1Shanghai Jiao Tong University, 2Hong Kong University of Science and Technology

This talk will presents a general measure for assessing the efficacy of control charts in detecting changes in the probability distribution of a sequence of observations. By establishing a dynamic, nonnegative, random threshold (control limit), we develop an approach for the construction of a generalized optimal control chart based on the general measure. This approach yields a unified formula of the minimal value of the measure for the generalized optimal chart. Special instances of the optimal control chart are examined, not only verifying the optimality of well-known the Shiryaev, Shiryaev-Roberts, Shewhart and CUSUM control charts in detecting the distribution change of the dependent observation sequence, but also obtaining the optimal modified EWMA control chart.

**Session 24CHI4: Advanced Statistical Learning For Complicated Gene Data**

**Large Precision Matrix Estimation for Compositional Data**

♦*Shucong Zhang*1, *Huiyuan Wang*2 *and Wei Lin*3.

1University of International Business and Economics， 2University of Pennsylvania， 3Peking University

High-dimensional compositional data are prevalent in many applications. The simplex constraint poses intrinsic challenges to inferring the conditional dependence relationships among the components forming a composition, as encoded by a large precision matrix. We introduce a precise specification of the compositional precision matrix and relate it to its basis counterpart, which is shown to be asymptotically identifiable under suitable sparsity assumptions. By exploiting this connection, we propose a composition adaptive regularized estimation (CARE) method for estimating the sparse basis precision matrix. We derive rates of convergence for the estimator and provide theoretical guarantees on support recovery and data-driven parameter tuning. Our theory reveals an intriguing trade-off between identification and estimation, thereby highlighting the blessing of dimensionality in compositional data analysis. In particular, in sufficiently high dimensions, the CARE estimator achieves minimax optimality and performs as well as if the basis were observed. We further discuss how our framework can be extended to handle data containing zeros, including sampling zeros and structural zeros. The advantages of CARE over existing methods are illustrated by simulation studies and an application to inferring microbial ecological networks in the human gut.

**Supervised Bayesian joint Graphical Model for Simultaneous Network Estimation and Subgroup Identification**

*Xing Qin*1, *Xu Liu*2, *Shuangge Ma*3 *and* ♦*Mengyun Wu*2.

1Shanghai University of International Business and Economics，2Shanghai University of Finance and Economics，3Yale School of Public Health

Heterogeneity is a fundamental characteristic of cancer. To accommodate heterogeneity, subgroup identification has been extensively studied and broadly categorized into unsupervised and supervised analysis. Compared to unsupervised analysis, supervised approaches potentially hold greater clinical implications. Under the unsupervised analysis framework, several methods focusing on network-based subgroup identification have been developed, offering more comprehensive insights than those restricted to mean, variance, and other simplistic distributions by incorporating the interconnections among variables. However, research on supervised network-based subgroup identification remains limited. In this study, we develop a novel supervised Bayesian graphical model for jointly identifying multiple heterogeneous networks and subgroups. In the proposed model, heterogeneity is not only reflected in molecular data but also associated with a clinical outcome, and a novel similarity prior is introduced to effectively accommodate similarities among the networks of different subgroups, significantly facilitating clinically meaningful biological network construction and subgroup identification. The consistency properties of the estimates are rigorously established, and an efficient algorithm is developed. Extensive simulation studies and a real-world application to TCGA data are conducted, which demonstrate the advantages of the proposed approach in terms of both subgroup and network identification.

**Mendelian Randomization Identifying and Estimating Causal Moderation Effect using Genome-Wide Summary Statistics**

♦*Xingjie Shi1*

*1*East China Normal University

Mendelian randomization (MR) employs genetic variants as instrumental variables to investigate the causal effect of an exposure on an outcome using summary statistics from genome-wide association studies. However, conventional MR approaches can be prone to omitted variable bias and model misspecification when the causal effect is influenced by an environmental variable. This study introduces MR-GEI, a novel method aimed at assessing the causal effect of exposure on outcome, while also examining how this effect is altered by an environmental factor. Through simulations, we show that MR-GEI is more effective in avoiding false positives resulting from moderation effects compared to existing methods. When applied to traits studied in recent GWAS research, MR-GEI reveals diverse causal relationships between males and females that are supported by existing literature, while also minimizing the identification of less plausible moderations.

**Joint Modeling Approach for Censored Predictors in Generalized Linear Model due to Detection Limit with Applications to Metabolites Data**

*Fengxue Li1, Shuo Bai1, Peng Ye2, Yi Tang3,* ♦*Hua He1*

1Tulane University, 2University of International Business and Economics, China,3University of Texas Health Science Center at Houston

Biomarker measurements obtained from urine, serum, or other biological matrices are commonly utilized in medical and health-related research. However, it is frequently observed that biomarker measures are left-censored due to their concentration levels falling below the limits of detection of the assay. When biomarker measurements are left-censored, they can originate from non-exposed subjects where their measures consistently register as zeros, thus being censored, or from exposed subjects whose exposure levels are below the detection limit threshold. In cases where censored biomarkers originate from both exposed and non-exposed subjects, the study population becomes mixed, valid inference is only assured if the mixed population is disentangled in data analysis. In this talk, we will present a joint modeling approach to handle censored predictors in generalized linear models, some simulation results, as well as findings on associations between metabolites and hypertension in the Bogalusa Heart Study.

**Session 24CHI124: Statistical Modeling of Functional Data and Related Applications in Biomedicine**

**Evaluating Prognostic Value of Dynamic of Circulating Lactate Dehydrogenase in Colorectal Cancer Using Modeling and Machine Learning**

♦*Haolun Ding1, Min Yuan2, Yaning Yang1, Manish Gupta3, Xu Steven Xu3*

1Department of Statistics and Finance, School of Management, University of Science and Technology of China， 2Department of Health Data Science, Anhui Medical University， 3Clinical Pharmacology and Quantitative Science, Genmab Inc., Princeton, New Jersey, USA

Pretreatment serum lactate dehydrogenase (LDH) levels have been associated with poor prognosis in several types of cancer, including metastatic colorectal cancer (mCRC). However, very few models link survival to longitudinal LDH measured repeatedly over time during treatment. We investigated the prognostic value of on-treatment LDH dynamics in mCRC. Using data from two large phase III studies (2L and 3L+ mCRC settings, n=824 and 210, respectively), we found that integrating longitudinal LDH data with baseline risk factors significantly improved survival prediction. Current LDH values performed best, enhancing discrimination ability (area under the receiver operating characteristic curve) by 4.5~15.4% and prediction accuracy (Brier score) by 3.9~15.0% compared with baseline variables. Combining all longitudinal LDH markers further improved predictive performance. After controlling for baseline covariates and other longitudinal LDH indicators, current LDH levels remained a significant risk factor in mCRC, increasing mortality risk by over 90% (P<0.001) in 2L patients and 60–70% (P<0.01) in 3L+ patients per unit increment in current log (LDH). Machine-learning techniques, like functional principal component analysis (FPCA), extracted informative features from longitudinal LDH data, capturing over 99% of variability and allowing prediction of survival. Unsupervised clustering based on the extracted FPCA features stratified patients into three groups with distinct LDH dynamics and survival outcomes. Hence, our approaches offer a valuable and cost-effective way for risk stratification and improves survival prediction in mCRC using LDH trajectories.

**A Nonparametric Concurrent Regression Model with Multivariate Functional Inputs**

*Yutong Zhai*1, ♦*Zhanfeng Wang*1, *Yuedong Wang*2

1University of Science and Technology of China， 2University of California - Santa Barbara

Regression models with functional responses and covariates have attracted extensive research. Nevertheless, there is no existing method for the situation where the functional covariates are bivariate functions with one of the variables in common with the response function. In this article, we propose a nonparametric function-on-function regression method. We construct model spaces using a Gaussian kernel function and smoothing spline ANOVA decomposition. We estimate the nonparametric function using penalized likelihood and study properties of the Gaussian kernel function and the convergence rate of the proposed estimation method. We evaluate the proposed methods using simulations and illustrate them using two real data examples.

**Statistical Inference for Large-scale Multi-source Heterogeneous Data**

♦*Jiuzhou Miao*1, *Li Cai*1, *Suojin Wang*2

1School of Statistics and Mathematics, Zhejiang Gongshang University， 2Department of Statistics, Texas A&M University.

In the era of digital information, people are faced with data that may not only be large-scale but also heterogeneous. In this paper, we study statistical inference for the overall population mean function of large-scale multi-source heterogeneous datasets. By borrowing hierarchical sampling methods and divide-and-conquer techniques, we propose a weighted local linear estimator for the overall population mean function of multi-source heterogeneous data. Through studying the pointwise convergence properties and extreme value distribution properties of the estimator, we construct asymptotically accurate simultaneous confidence bands and pointwise confidence intervals for large-scale multi-source heterogeneous data. Our proposed methods are applicable not only to scenarios of heterogeneous data but also to scenarios of homogeneous data using divide-and-conquer methods. Numerical simulation studies show that the proposed methods perform well in analyzing both large-scale multi-source heterogeneous data and homogeneous data. As an illustration, we apply the proposed methods to hypothesis testing problems on Beijing multi-site air-quality data and cardiovascular research data.

**Application of Multi-Criteria Decision Analysis (MCDA) to Develop an Efficacy-Safety Profile for Dose Selection in Oncology/Hematology Trials**

♦*Chengqing Wu*1, *Jun Zhang*1, *Hao Wang*2

1BeiGene， 2Iowa State University

In the recent oncology/hematology drug development, dose optimization and selection has been a challenge and there is a strong business need to have a consistent and transparent framework to facilitate the dose selection decision making especially in the early phase of the drug development. One major challenge in dose selection is how to incorporate the totality of the data (e.g., efficacy and safety) into the decision making. Multi-Criteria Decision Analysis (MCDA) is a methodology for supporting decision making when multiple objectives need to be assessed. MCDA method can support decision makers faced with evaluating alternatives by taking into account multiple criteria in an explicit manner. In this project, we developed a MCDA framework to assess the efficacy-safety profile for dose selection and study the applicability of the framework in the settings of oncology/hematology trials. Due to the nature of the MCDA, extensive simulation studies are conducted to assess the operating characteristics of the proposed framework across different settings (e.g., weighting schedules, sample size, number of doses, etc.).

**Session 24CHI101: Semiparametric Panel Data Models: Theory and Applications**

**Robust Inference for High-Dimensional Panel Data Models**

*Jiti Gao*1, *Bin Peng*1 *and* ♦*Yayi Yan*2.

1Monash University， 2Shanghai University of Finance and Economics

In this paper, we consider estimation and robust inference for high-dimensional panel data models, where the number of regressors can grow faster than the sample size. We extend several Lasso techniques to a panel data setting allowing for non-Gaussian, serially and cross-sectionally correlated and heteroskedastic error processes for the purposes of consistent estimation, variable selection and asymptotic normality respectively. For the construction of confidence intervals, we further consider the estimation of a high-dimensional long-run covariance matrix using a thresholded estimator. We obtain a sharp convergence rate for the thresholded estimator by developing a new Nagaev-type concentration inequality for quadratic forms of panel data. We also demonstrate the usefulness of these theoretical results by investigating a high-dimensional panel data model with interactive fixed effects. To validate the theoretical results, we conduct extensive simulations. Finally, we demonstrate the empirical relevance by applying the framework to a high-dimensional asset pricing model of stock returns.

**Macroeconomic Factors and Return Predictability: A Factor-Augmented Sum-of-the-Parts Method**

*Tingting Cheng*1, ♦*Xuanbin Yang*1 *and Bo* *Zhao*1

1Nankai University

In this paper, we augment the sum-of-the-parts (SOP) method in Ferreira and Santa-Clara (2011) with estimated factors from a high-dimensional macroeconomic dataset to forecast the stock market returns, namely the factor-augmented sum-of-the-parts (FA-SOP) method. Our approach capitalizes on the disparate time series persistence observed among return components and leverages the inherent strengths of macroeconomic factors in characterizing the state of economy. With monthly data, our model achieves out-of-sample R-squares surpassing 1.9% and reaches 6.0% with quarterly data, which notably surpasses the predictive regression, factor- augmented regression, and SOP method, demonstrating superior statistical and economic gains. Additionally, we propose a present-value model integrating macroeconomic factors to elucidate the mechanism underlying the determination of return components. A Monte Carlo simulation study is conducted to demonstrate the efficacy of our model in explaining the performance variations among different methodologies when applied to real data.

**Estimation and Inference for 3-Dimensional Panel Data Models**

*Guohua Feng*1, *Jiti Gao*2,♦*Fei Liu*3 *and Bin Peng*2.

1University of North Texas， 2Monash University， 3Nankai University

In this paper, we provide estimation and inference for 3-dimensional panel data models with heterogeneous coefficients. Our 3-dimensional panel data models specify the nature of common shocks through the use of a hierarchical factor structure (i.e., global factors and sector factors). Accordingly, we propose an approach to estimating the hierarchy, thus enabling us to have a better understanding of the relative importance of the two types of unobservable shocks. Second, we provide bias-corrected estimators and give bootstrap procedures to construct the confidence intervals for the parameters of interest while allowing for correlation along three dimensions of idiosyncratic errors. We justify the theoretical findings using extensive simulations. In the empirical study, we examine the twin hypotheses of conditional and unconditional convergence for manufacturing industries across countries.

**Semiparametric Estimation of Heterogeneous Treatment Effects with Sieve Method**

♦*Jichang Yu*1

1Zhongnan University of Economics and Law

**Session 24CHI98: Reinforcement Learning and Personalized Medicine**

Identifying heterogeneous treatment effects in observational studies is very difficult due to the fact that the outcome model or the treatment assignment model must be correctly specified. Taking advantages of semiparametric model, we use the single-index model to estimate heterogeneous treatment effects, which can allow the link function to be unbounded and have unbounded support. The link function is regarded as a point in an infinitely dimensional function space, and we can estimate the link function and the index parameter simultaneously. We establish the asymptotic properties of the proposed estimator. The finite-sample performance of the proposed estimator is evaluated through simulation studies. The proposed method is illustrated on real data from Pennsylvania in the USA to investigate the effect of maternal smoking during pregnancy on birth weight.

**Online Statistical Inference for Low-rank Reinforcement Learning**

♦*Wei Sun*1

1Purdue University

The study of online decision-making problems that leverage contextual information has drawn notable attention due to their significant applications in fields ranging from healthcare to autonomous systems. In modern applications, such context format can be rich and can often be formulated as higher-order tensors. Moreover, while existing online decision algorithms mainly focus on reward maximization, less attention has been paid to statistical inference. To fill in these gaps, in this work, we consider an online decision-making problem with high-order contextual information where the true model parameters have a low-rank structure. We propose a fully online procedure to conduct statistical inference with adaptively collected data. The low-rank structure of the model parameter and the adaptivity nature of the data collection process make this difficult: standard low-rank estimators are biased and cannot be obtained in a sequential manner while existing inference approaches in sequential decision-making algorithms fail to account for the low-rankness and are also biased. To address these, we introduce a new online debiasing procedure to simultaneously handle both sources of bias.

**Off-policy Evaluation in Doubly Inhomogeneous Environments**

♦*Zeyu Bian*1, *Chengchun Shi*2, *Zhengling Qi*3 *and Lan Wang*1.

1University of Miami， 2London School of Economics and Political Science， 3George Washington University

This work aims to study off-policy evaluation (OPE) under scenarios where two key reinforcement learning (RL) assumptions -- temporal stationarity and individual homogeneity are both violated. To handle the ``double inhomogeneities", we propose a class of latent factor models for the reward and observation transition functions, under which we develop a general OPE framework that consists of both model-based and model-free approaches. We establish the theoretical properties of the proposed value estimators and empirically show that our approach outperforms competing methods that ignore either temporal nonstationarity or individual heterogeneity. Finally, we illustrate our method on a data set from the Medical Information Mart for Intensive Care.

**Policy Learning with Continuous Actions Under Unmeasured Confounding**

*Yuhan Li*1, *Eugene Han*1, *Wenzhuo Zhou*2 , *Zhengling Qi*3 , Yifan Cui4 *and* ♦*Ruoqing Zhu*1.

1University of Illinois Urbana Champaign， 2University of California, Irvine， 3George Washington University 4Zhejing University

In the realm of reinforcement learning for personalized medicine, the presence of unmeasured confounding variables often impedes the learning of optimal treatment policies. This study aims to address the challenge of policy learning in environments with continuous actions under unmeasured confounding. We introduce a novel statistical framework that leverages advanced techniques in sufficient dimension reduction and causal inference to account for latent confounders. By employing computationally efficient algorithms designed for high-dimensional data, our approach achieves improved policy learning outcomes while maintaining computational tractability. Extensive simulations and a case study in sepsis treatment validate the efficacy of the proposed method in improving decision-making quality. Our results shed new light on the robustness and effectiveness of continuous action policy learning under the complex setting of unmeasured confounding.

**Proxy-aided Demand Learning with an Application on Various Pricing Problems**

♦*Tao Shen*1 *and Yifan Cui*2

1National University of Singapore，2Zhejiang University

In data-driven demand learning, understanding customer willingness to pay presents a significant challenge due to the complex interplay between various influencing factors. This paper addresses the multifaceted relationship between quantities like price and sales, highlighting the difficulties in identifying the causal effect with the existence of unmeasured confounders. To mitigate bias in evaluating pricing decisions, we introduce proxy variables into the demand learning process. The paper further explores data-driven pricing challenges within a confounded environment, focusing on static pricing and personalized pricing problems, showcasing the practical application of the proposed demand learning process.

**Session 24CHI8: Advancements in Statistical Methods for Diverse Applications**

**Demographic Parity-aware Individualized Treatment Rules**

*Wenhai Cui*1, ♦*Wen Su*2, *Xiaodong Yan*1 *and Xingqiu Zhao*3.

1Shandong University， 2City University of Hong Kong， 3The Hong Kong Polytechnic University

There has been a growing interest in the development of advanced methodologies aimed at estimating the optimal individualized treatment rules (ITR) in various fields, such as business decision making, precision medicine, and social welfare distribution. The application of individualized treatment rules (ITR) within a societal context raises substantial concerns regarding the potential for unintended discrimination. Customized policies learned from biased data can inadvertently lead to disparities based on sensitive attributes such as age, gender, or race. To address this concern directly, we introduce a tailored nonlinear fairness constraint that aligns with the requirements of demographic parity (DP) ITR. We obtain the optimal demographic parity-aware ITR solution by solving a non-convex constrained optimization problem. To overcome computational challenges, we identify linear and nonlinear fairness constraint proxies and leverage the support vector machine framework to transform it into a convex quadratic programming problem. Additionally, we establish the asymptotic consistency and convergence rate of the proposed estimator. We demonstrate performance of the proposed method through extensive simulation studies and a real data analysis utilizing the entrepreneurial program data.

**Diverging-Dimensional Distributed Estimation under GLMs with Pilot Samples**

♦*Xiaochao Xia*1

1Chongqing University

Statistical inference is often of primary interest in the analysis of distributed data. However, most of the existing works focus on the situation where the dimension $p$ is fixed. Less has considered statistical inference on the diverging dimensional distributed estimation. In this paper, we investigate a distributed framework based on a pilot sampling procedure to address the challenges of non-randomly distributed data. We first propose two distributed estimators based on pilot sampling procedure under generalized linear models, which are communication-efficient and require to solve distributed optimizations for non-randomly distributed big data. Two concrete algorithms are provided afterwards. More importantly, we systematically study the theoretical properties of our estimators, including the rate of convergence and asymptotic distributions in the diverging dimensional case $p<n$, as well as the non-asymptotic error bounds of the proposed regularization methods in the high dimensional case $p>n$. Under regularity conditions, theoretical results show that the proposed estimators can achieve the same efficiency as the global estimators. Finally, our proposed methods are evaluated through numerical examples.

**A New Model-Free Feature Screening Procedure for Ultrahigh-Dimensional Interval-Censored Failure Time Data**

♦*Jing Zhang*1, *Mingyue Du*2, *Yanyan Liu*3 *and Jianguo Sun*4

1Zhongnan University of Economics and Law, 2The Hong Kong Polytechnic University, 3Wuhan University and 4University of Missouri

Screening important features based on ultrahigh-dimensional data has become an important task in statistical analysis. As such, several screening procedures have been proposed for various types of studies or data, including complete data and right-censored failure time data. In this study, we consider ultrahigh-dimensional interval-censored failure time data. Such data occur frequently in medical follow-up studies, among others, and include right-censored data as a special case, but for which few works exist. For the problem, a distance correlation-based sure independent screening procedure is proposed. The new approach is model-free and does not require estimating survival functions, unlike most existing nonparametric screening procedures for failure time data. We establish the sure screening property and the ranking consistency of the proposed method, and conduct an extensive simulation study, which suggests that the proposed procedure works well for practical situations. Finally, we apply the proposed method to a set of real data on Alzheimer’s disease, which motivated this study.

**Over-parameterized Deep Nonparametric Regression for Dependent Data with Its Applications to Reinforcement Learning**

*Xingdong Feng*1, *Yuling Jiao*2,♦*Lican Kang*2, *Baqun Zhang*1, *Fan Zhou*1

1Shanghai University of Finance and Economics, 2Wuhan University

In this paper, we provide statistical guarantees for over-parameterized deep nonparametric regression in the presence of dependent data. By decomposing the error, we establish nonasymptotic error bounds for deep estimation, which is achieved by effectively balancing

the approximation and generalization errors. We have derived an approximation result for Holder functions with constrained weights. Additionally, the generalization error is bounded by the weight norm, allowing for a neural network parameter number that is

much larger than the training sample size. Furthermore, we address the issue of the curse of dimensionality by assuming that the samples originate from distributions with low intrinsic dimensions. Under this assumption, we are able to overcome the challenges posed by

high-dimensional spaces. By incorporating an additional error propagation mechanism, we derive oracle inequalities for the over-parameterized deep fitted Q-iteration.

**Session 24CHI21: Causal Inference and Semiparametric Learning**

**Root-n Consistent Semiparametric Learning with High-Dimensional Nuisance Functions under Minimal Sparsity**

*Lin Liu1* and♦*Yuhao Wang2.*

1Shanghai Jiao Tong University， 2Tsinghua University

Treatment effect estimation under unconfoundedness is a fundamental task in causal inference. In response to the challenge of analyzing high-dimensional datasets collected in substantive fields such as epidemiology, genetics, economics, and social sciences, many methods for treatment effect estimation with high-dimensional nuisance parameters (the outcome regression and the propensity score) have been developed in recent years. However, it is still unclear what is the necessary and sufficient sparsity condition on the nuisance parameters for the treatment effect to be \sqrt{n}-estimable. In this paper, we propose a new Double-Calibration strategy that corrects the estimation bias of the nuisance parameter estimates computed by regularized high-dimensional techniques and demonstrate that the corresponding Doubly-Calibrated estimator achieves 1 / \sqrt{n} -rate as long as one of the nuisance parameters is sparse with sparsity below \sqrt{n} / \log p, where p denotes the ambient dimension of the covariates, whereas the other nuisance parameter can be arbitrarily complex and completely misspecified. The Double-Calibration strategy can also be applied to settings other than treatment effect estimation, e.g. regression coefficient estimation in the presence of diverging number of controls in a semiparametric partially linear model.

**A Semiparametric Instrumented Difference-in-Differences Approach to Policy Learning**

♦*Pan Zhao1* and *Yifan Cui2.*

1Inria，2Zhejiang University

Recently, there has been a surge in methodological development for the difference-in-differences (DiD) approach to evaluate causal effects. Standard methods in the literature rely on the parallel trends assumption to identify the average treatment effect on the treated. However, the parallel trends assumption may be violated in the presence of unmeasured confounding, and the average treatment effect on the treated may not be useful in learning a treatment assignment policy for the entire population. In this article, we propose a general instrumented DiD approach for learning the optimal treatment policy. Specifically, we establish identification results using a binary instrumental variable (IV) when the parallel trends assumption fails to hold. Additionally, we construct a Wald estimator, novel inverse probability weighting (IPW) estimators, and a class of semiparametric efficient and multiply robust estimators, with theoretical guarantees on consistency and asymptotic normality, even when relying on flexible machine learning algorithms for nuisance parameters estimation. Furthermore, we extend the instrumented DiD to the panel data setting. We evaluate our methods in extensive simulations and a real data application.

**Causal Inference for Dyadic Data with Interference**

♦*Wang Miao*1.

1Peking University

Estimating the global average treatment effect on a network could be considerably biased due to spillover effects in the presence of unknown network interference. We consider novel dyadic outcomes in the presence of interference. Such outcomes are common in many social network sources, such as forwarding a message or sharing a link. We first establish the causal inference framework for dyadic outcome in network interference, which is of particular interest in online experimentation in many social media companies. Then we show that the unbiased estimator for the global average treatment effect based on the conventional outcomes does not exist in most cases. We provide subsequently unbiased estimators based on dyadic outcomes for randomized experiments. We derive the variance bound of the proposed estimators and provide a variance estimator for quantifying the estimation uncertainty. We also illustrate with a variety of numerical experiments and an application about an online experiment in Wechat Channels.

**Distribution-Free Prediction Intervals Under Covariate Shift, With an Application to Causal Inference**

*Jing Qin, ♦Yukun Liu, Moming Li and Chiung-Yu Huang*

Owing to its appealing distribution-free feature, conformal inference has become a popular tool for constructing prediction intervals with a desired coverage rate. However, most existing methods rely on data-splitting techniques, which often result in longer intervals and unreliable coverage rates with a finite sample size in the presence of covariate shift. To address these challenges, we propose methods based on a pivotal quantity derived under a parametric working model and employ a resampling-based framework to approximate its distribution. The resampling-based approach can produce prediction intervals with a desired coverage rate without splitting the data and can be easily applied to causal inference settings where a shift in the covariate distribution can occur between treatment and control arms. Additionally, the proposed approaches enjoy a double robustness property and are adaptable to different prediction tasks. Our extensive numerical experiments demonstrate that, compared to existing methods, the proposed novel approaches can produce substantially shorter conformal prediction intervals with lower variability in the interval lengths while maintaining promising coverage rates and advantage in versatile usage.

**Session 24CHI58: Real World Evidence: Under the Regulatory Microscope**

**The Hainan RWE Pathway**

*♦Sheng Feng1*

1Parexel

**Real World Evidence-Under the Regulatory Microscope**

♦*Feng Sun*.

**Integrative Analysis of Randomized Clinical Trials with Real-World Data: Methods and Applications**

♦*Xiaofei Wang1* and *Shu Yang1.*

1Duke University 2North Carolina State University

In clinical research, individual patient data from both randomized trials and large observational studies on the same disease and treatments exist. Therefore, there is a great interest and need to integrate the data from these multiple data sources. Randomized clinical trials (RCTs) offer the highest level of evidence of treatment safety and efficacy as randomization eliminates both measured and unmeasured confounders. Patients enrolled in randomized trials are conveniently ascertained and represent a more restrictive patient group of the target real-world patient population to which the new treatment will be given. The estimated treatment effects based on only the RCT data using the standard methods often lack precision and external validity for the target population. On the other hand, real-world or observational studies often contain a much larger number of patients of the same disease and represent either a random sample of the target population. Due to lack of treatment randomization, there are always concerns over measured and unmeasured confounders. Given the complementarity of RCTs and observational studies, integrated analysis approaches are called for to efficiently exploit the relative strengths of the data from both RCTs and observational studies and carefully address their potential drawbacks. In this talk, we first discuss the common questions that investigators often have when planning an integrated analysis with parallel data from RCTs and observational studies. We then review several recent methods that exploit the complementary features of RCTs and observational studies. These methods allow us to answer these questions by providing robust and efficient estimates of the average treatment effect (ATE) with external validity and by offering test-then-pool elastic method and bias modelling method to address both measured and unmeasured confounding in the integrative analysis. These methods will also be illustrated with real data examples.

**Single-Arm Trials with External Controls**

♦*Jie Chen*

Single-arm trials (SATs) are suitable in some specific scenarios for accelerated drug approval, in which external controls are often used to estimate comparative effectiveness. This talk first describes conditions under which SATs are appropriate and then types of external controls and their respective pros and cons. To obtain more accurate and precise estimate of treatment effect, the targeted-learning framework is applied by (1) starting from defining the esitmand, (2) describing observed data distributions, (3) specifying causal estimand, (4) exploring the causal identifiability, (5) estimating the statistical and causal estimands, and (6) sensitivity analysis. An example is given to illustrate how the TL-based framework can be used in SATs.

**Session 24CHI125: New developments in health data analysis with applications to clinical trials**

**Policy Learning with Distributional Welfare**

♦*Yifan Cui1 and Sukjin Han2*

1Zhejiang University，2University of Bristol

In this paper, we explore optimal treatment allocation policies that target distributional welfare. Most literature on treatment choice has considered utilitarian welfare based on the conditional average treatment effect (ATE). While average welfare is intuitive, it may yield undesirable allocations especially when individuals are heterogeneous (e.g., with outliers) - the very reason individualized treatments were introduced in the first place. This observation motivates us to propose an optimal policy that allocates the treatment based on the conditional quantile of individual treatment effects (QoTE). Depending on the choice of the quantile probability, this criterion can accommodate a policymaker who is either prudent or negligent. The challenge of identifying the QoTE lies in its requirement for knowledge of the joint distribution of the counterfactual outcomes, which is generally hard to recover even with experimental data. Therefore, we introduce minimax policies that are robust to model uncertainty. A range of identifying assumptions can be used to yield more informative policies. For both stochastic and deterministic policies, we establish the asymptotic bound on the regret of implementing the proposed policies. In simulations and two empirical applications, we compare optimal decisions based on the QoTE with decisions based on other criteria. The framework can be generalized to any setting where welfare is defined as a functional of the joint distribution of the potential outcomes.

**Regularized T Distribution: Definition, Properties and Applications**

*Zongliang Hu1, Yiping Yang2, Gaorong Li 3and* ♦*Tiejun Tong4*

1Shenzhen University， 2Chongqing Technology and Business University， 3Beijing Normal University， 4Hong Kong Baptist University

For gene expression data analysis, an important task is to identify genes that are differentially expressed between two or more groups. Nevertheless, as biological experiments are often measured with a relatively small number of samples, how to accurately estimate the variances of gene expression becomes a challenging issue. To tackle this problem, we introduce a regularized t distribution and derive its statistical properties including the probability density function and the moment generating function. The noncentral regularized t distribution is also introduced for computing the statistical power of hypothesis testing. For practical applications, we apply the regularized t distribution to establish the null distribution of the regularized statistic, and then formulate it as a regularized t-test for detecting the differentially expressed genes. Simulation studies and real data analysis show that our regularized t-test performs much better than the Bayesian t-test in the “limma” package, in particular when the sample sizes are small.

**Variable Selection for Interval-censored Failure Time Data**

*(Tony) Jianguo Sun*

Interval-censored failure time data occur in many areas,

including demographical studies, economic studies, medical studies and social sciences, and in different forms. This talk will discuss variable selection for such data and present some recently developed tools.

**Bayesian Sample Size Estimation in Pediatric Clinical Trials with Adult Extrapolation**

*Fangrong Yan* *and* ♦*Ruyue He*.

Pediatric drug development has always been an important challenge in the development of new drugs, mainly because of small sample sizes due to difficulties in patient recruitment, which often makes it difficult to attain the statistical significance and statistical power required for clinical trials. Employing Bayesian information borrowing in clinical trial design is an important approach to address this issue, which has gained consistent recognition from regulatory bodies. Within the framework of Bayesian information borrowing-related trial designs, sample size estimation is a critical issue. According to the relevant guidance principles issued by NMPA, estimating pediatric sample sizes typically involves subtracting the prior effective sample size from an estimated maximum sample size. However, this method does not directly calculate sample sizes, and the differences in ESS calculation methods can lead to unreliable results. Therefore, it is urgent to develop a scientific, reasonable and reliable sample size calculation method. We propose a method for calculating sample sizes based on the highest density posterior interval of the posterior distribution, while ensuring type I errors and power.

**Session 24CHI96: Recent Statistical Methods and Applications in Precision Medicine**

1. **value: a superior alternative to P-value and its adjustments in DNA methylation studies**

♦*Xiaoqing Pan1, Yifan Yang1* and *Pengyuan Liu2 .*

1Shanghai Normal University， 2Zhejiang University

DNA methylation plays a crucial role in transcriptional regulation. Reduced representation bisulfite sequencing (RRBS) is a technique of increasing use for analyzing genome-wide methylation profiles. Many computational tools such as Metilene, MethylKit, BiSeq and DMRfinder have been developed to use RRBS data for the detection of the differentially methylated regions (DMRs) involved in epigenetic regulations of gene expression. For DMR detection tools, as for countless other medical applications, P-values and their adjustments are among the most standard reporting statistics used to assess the statistical significance of biological findings. However, P-values are coming under increasing criticism relating to their questionable accuracy and relatively high levels of false positive or negative indications. In this talk, I will introduce our method and R package ‘metevalue’ to calculate E-values, as likelihood ratios falling into the null hypothesis over the entire parameter space, for DMR detection in RRBS data. To evaluate the performance of E-values, we generated various RRBS benchmarking datasets using our simulator ‘RRBSsim’ with 8 samples in each experimental group. Our comprehensive benchmarking analyses showed that using E-values not only significantly improved accuracy, AUC and power, over that of P-values or adjusted P-values, but also reduced false discovery rates and type I errors. In applications using real RRBS data of CRL rats and a clinical trial on low-salt diet, the use of E-values detected biologically more relevant DMRs and also improved the negative association between DNA methylation and gene expression.

**Optimal Treatment Decision Making with Multiple Treatments under Proportional Hazards Model**

♦*Yuexin Fang, Xiangyong Tan* and *Qian Li.*

Under multiple treatment options, a new proportional hazards model with an unknown baseline covariate effect and interaction between the treatment and covariates on censored survival data is considered. We use the multivariate logistic regression to model the propensity score, and construct the doubly robust estimating equations based on A-learning method and the time-dependent propensity score. The asymptotic properties of the proposed estimators are established when either the baseline effect model for covariates or the propensity score is correctly specified. Numerical simulations and an application to AIDS clinical trial data set illustrate the proposed method.

**Global Identification and Characterization of LncRNAs as Transcriptomic Key Regulators in Human Cancer**

♦*Pengyuan Liu1, Chuanzheng Lup1, Xiang Cheng1, Liyuan Zhou1, Xiaoqing Pan2*.

1Zhejiang University， 2Shanghai Normal University

Background:

The mammalian transcriptional regulatory network orchestrates various biological processes, with key drivers often dictating system-wide state changes. While mRNA molecules have been extensively studied as network drivers, the potential role of non-coding RNAs, particularly Long non-coding RNAs (LncRNAs), remains relatively unexplored. With emerging evidence suggesting the functional significance of LncRNAs, this study aims to investigate their potential as pivotal contributors to transcriptome programs and elucidate their underlying biological characteristics as network drivers.

Methods:

Building upon our previous work, we introduce an enhanced computational framework, WMDS.net, tailored to identify key driver LncRNAs. Leveraging expression differences across physiological states, we construct a co-expression network, where edge weights are determined by integrating node degrees and the significance of gene co-expression relationships across states. This approach quantifies the contribution of each node to the transcriptional regulatory network. Subsequently, we conduct a comprehensive analysis of the identified driver LncRNAs, considering sequence conservation, tissue specificity, transcript length, expression patterns, and transcript types, among other characteristics.

Results:

Applying the enhanced WMDS.net model to 14 cancer RNA-seq datasets from TCGA, we identify driver LncRNAs associated with specific cancer types and pan-cancer drivers. Our findings reveal significant differences in LncRNA-related characteristics between pan-cancer drivers and other LncRNAs, including sequence conservation, transcript length, and exon number. Furthermore, utilizing pan-cancer drivers as features, we develop high-performance machine learning diagnostic models for certain cancers and achieve classification of tumor subtypes, demonstrating notable differences in patient survival rates.

Conclusions:

In summary, our study sheds light on the pivotal role of LncRNAs as key players in the transcriptional network. The distinctive biological characteristics of identified pan-cancer drivers underscore their critical roles within the network, providing valuable insights for understanding transcriptome regulation and cancer pathogenesis.

**The theory of sieve method based on Laguerre polynomial system and its applications**

♦*Xiumei Reng*

This report focuses on the sieve method for nonnegative stationary data. By applying the Laguerre polynomials for orthogonal expansion of unknown object functions, we established a new sieve method and the central limit theorem of related estimators. Compared with the existing theoretical framework such as the sieve method based on the Hermit orthogonal system, the sieve method we established has a wider searching scope of the objective functions in the model with non-negative stationary data, and owns a potential application prospect in the specific applications related to interest rates, stock trading volume, input and output indicators, and green financial development index.

**Session 24CHI106: Special Considerations in Oncology Drug Development**

**Statistical consideration in backfill cohort approaches**

♦*Michael Lee.*1

Traditionally, the primary objective in oncology dose-escalation studies was to detect the most tolerable dose (MTD) for an acceptable toxicity level. Since the immunotherapy era, the desire therapeutically efficacious dose could be much lower than the MTD. Therefore, the primary objective of a dose-escalation study could be to determine a recommended expansion dose (RED). After the FDA Project Optimus, in addition to the RED determination objective, more and more studies focus on efficacy exploration and the RED determination weighs in efficacy evidence. This is one of reasons that backfill cohort approaches become popular recently. A dose-escalation study with backfill cohorts can increase the efficiency in drug development process, however, a backfill cohort design will also increase study cost and prolong study duration. There must be a sweet point between the size and timing of backfill cohort to maximize the study value. In this presentation we will use numerical examples to illustrate general considerations in backfill cohort.

**The application of Restricted Mean Duration of Response (RMDoR)**

♦*Dian Yang*

Response-based endpoints such as time to response, objective response, and duration of response (DOR) are frequently employed in oncology clinical trials to ascertain therapeutic effects, thereby aiding in proof-of-concept or submission decisions. A recent proposition was the restricted mean DOR (RMDOR), a composite nonparametric method designed to effectively measure the treatment effect associated with tumor reductions. This method provides a straightforward approach to conduct statistical inference in cross-arm comparison and has been incorporated in several Phase III trials..

**Key Considerations of Surrogate Endpoints in Oncology Clinical Trials**

♦*Nan Li1*

1MSD China

Though overall survival (OS) is the accepted golden rule in evaluating oncology clinical trials, established or validated surrogate endpoints are often used in traditional or accelerated approval by health authorities when considering serious or life-threatening illnesses with limited availability of alternative treatment options. Recently, Food and Drug Administration (FDA) held an Oncologic Drugs Advisory Committee (ODAC) meeting decided that evidence supported minimal residual disease (MRD) could be used as accelerated approval endpoint in clinical trials of multiple myeloma. This is historical in accepting MRD-negative complete remission (CR) as a reliable early endpoint. Reviewing related guidelines, we summarize some key considerations when discussing new surrogate endpoint with-regulatory. Some specific novel endpoint development is presented.

**Consideration of stratification in oncology confirmatory trials with time-to event Endpoint**

♦*Zhiji Tang , Zhaohua Lu , Yuan Geng* and *Philipe He* .

Stratification in randomization and analysis is widely applied in confirmatory oncology clinical trials to balance treatment group. Stratified analysis is commonly used to control confounding factors in evaluating treatment effect at both design and analysis stages. However, under-stratification and over-stratification potential leading a power loss is an interesting topic in the typical setting of a confirmatory clinical trial. We reviewed some current papers about consideration of stratification impact for the power and the method of determine number strata to avoid over-stratification. In these papers, extensive simulation studies to evaluate the impact of under-stratification over-stratification or mis-stratification on the power with overall study result of survival analysis and the estimate of hazard ratio using stratified log-rank test and Cox PH model, respectively. The difference in power between stratified and unstratified log-rank tests is also investigated under different scenarios. Additionally, methods of combining small strata are explored and compared and a method is proposed to restrict the number of strata to reduce possibility of over-stratification. In addition, we proposed a framework to evaluate the impact of stratification on the probability of showing consistency in MRCT clinical trials.

**Session 24CHI87: Recent Developments for Analyzing Complex Heterogeneous Data**

**Penalized Robust Regression Estimator via Discounted Exponentile Loss for High-dimensional Heterogeneous Data**

♦*Yunlu Jiang* 1*, Zou Hang*1*, Wen Canhong*2*, Zhang Baoxue3, Wang Xueqin2*

1Jinan University， 2University of Science and Technology of China， 3Capital University of Economics and Business

High-dimensional data in biomedicine, econometrics, and finance often exhibit heteroscedasticity, attracting significant attention in research. While numerous methods have been proposed to address heteroscedasticity or heavy-tailed errors, many of these methods lack theoretical robustness characterization and are not robust against high leverage points. To overcome these limitations, this paper presents a novel robust variable selection procedure for high-dimensional heterogeneous data. Our approach introduces an asymmetric exponential squared loss function and achieves the highest asymptotic breakdown point under mild conditions. Furthermore, it ensures variable selection consistency and asymptotic normality. Empirical results demonstrate the superior performance of our proposed method across various settings, particularly in the presence of high leverage points in high-dimensional, heavy-tailed, and heterogeneous datasets.

**Penalized Estimation and Order Selection for Hidden Markov Models**

♦*Mian Huang*1, *Yong Lin*1, *Weixin Yao*2 1Shanghai University of Finance and Economics， 2UC Riverside

Determining the unknown number of hidden states is a key challenge in the estimation of Hidden Markov models (HMMs). In this paper, we propose a new penalized likelihood approach for estimating the number of states in an over fitted HMM. The method introduces two penalties on the likelihood - one ensures the state distributions are bounded away from each other, and the other penalizes the column-wise transition probabilities to eliminate redundant states. Under certain regularity conditions, we show that the resulting penalized estimator consistently estimates the true number of states, and the parameter estimation is asymptotically normal. An efficient EM algorithm is provided to obtain the penalized maximum likelihood estimates. Simulation studies and an application demonstrate finite sample performance of the proposed method.

**Online Change-point Detection for Functional Data**

*Hanbing Zhu*1, ♦*Houlin Zhu*1, *Yehua Li*2, *Xuejun Wang*1, *Xinyuan Song*3

1Anhui University，2Department of Statistics, University of California, Riverside， 3The Chinese University of Hong Kong

We propose a CUSUM-type online change-point detection procedure for monitoring a change in the mean function of dependent functional

data. Our method is fully nonparametric and does not require dimension reduction for functional observations. For the proposed sequential monitoring scheme we provide the limiting distribution under the null hypothesis of no change which yields the threshold to control the global false alarm rate asymptotically. Furthermore, we show that the proposed sequential test has asymptotic power one. The method is investigated by means of Monte Carlo simulation studies and an application to a real data set.

**New Regression Model: Modal Regression**

♦*Weixin Yao.*

Built on the ideas of mean and quantile, mean regression and quantile regression are extensively investigated and popularly used to model the relationship between a dependent variable Y and covariates x. However, the research about the regression model built on the mode is rather limited. In this talk, we propose a new regression tool, named modal regression, that aims to find the most probable conditional value (mode) of a dependent variable Y given covariates x rather than the mean that is used by the traditional mean regression. The modal regression can reveal new interesting data structure that is possibly missed by the conditional mean or quantiles. In addition, modal regression is resistant to outliers and heavy-tailed data, and can provide shorter prediction intervals when the data are skewed. Furthermore, unlike traditional mean regression, the modal regression can be directly applied to the truncated data. Modal regression could be a potentially very useful regression tool that can complement the traditional mean and quantile regressions.

**Session 24CHI118: Statistical Learning with Multiple Data Sources**

**Communication Efficient Federated Statistical Analysis with Differential Privacy**

*Changgee Chang1, Zhiqi Bu2* and♦*Qi Long3*

1Indiana University School of Medicine， 2Amazon AWS AI， 3University of Pennsylvania

Electronic health records (EHRs) offer great promises for advancing precision medicine and, at the same time, present significant analytical challenges. Particularly, it is often the case that patient-level data in EHRs cannot be shared across institutions (data sources) due to government regulations and/or institutional policies. As a result, there are growing interests about distributed learning over multiple EHRs databases without sharing patient-level data. To tackle such challenges, we propose a novel communication efficient federated statistical analysis approach that aggregates the local optimal estimates, by turning the problem into a missing data problem. In addition, we propose incorporating posterior samples of remote sites, which can provide partial information on the missing quantities and improve efficiency of parameter estimates while having the differential privacy property and thus reducing the risk of re-identification and information leaking. The proposed approach, without sharing the raw patient level data, allows for proper statistical inference and can accommodate sparse regressions. We also provide theoretical investigation for the asymptotic properties of the proposed method for statistical inference and differential privacy. We demonstrate the superiority of our proposed method over several existing methods in extensive simulations and analysis of EHRderived data from multiple hospitals in the Georgia Coverdell Acute Stroke Registry.

**Individualized and adaptive transfer learning via subset-estimating equation and composition**

*Lu Lin1, Jun Lu2* and♦*Xiaochen Zhang3* .

1Shandong University， 2National University of Defense Technology 3Beijing Normal University

The common methods of transfer learning ignore, to some extent, the individual difference among the source models, and rely heavily on the similarity condition between the source models and the target model. In this paper, an individualized and adaptive transfer learning is proposed through a new statistical strategy: subset-average estimating equation of multi-task learning. According to the individual and common features of all the source models, the differences between the source models and the target model are decomposed into the individual and common deviations. Correspondingly, the subset-average estimating equations for the individual and common deviations are defined, and the transfer learning estimators of the target parameters are then obtained. Such a strategy is adaptive to the individual and common features of the multi-source models, and is also adaptive to the distance between the source models and the target model. By the optimal individual tuning parameters, the convergence rate of the transfer learning estimation can be improved if the similarity condition holds, and the estimation efficiency can be enhanced if without the similarity condition. Furthermore, the method of subset-average estimating equation has potential applications in other scenarios of heterogenous datasets, such as clustering, change-point analysis and meta analysis. The comprehensive simulation studies and real data analysis can illustrate that the new strategy is easy to use, the resulting transfer learning is significantly better than the competitors, and is comparable with the oracle versions.

**Information-Incorporated Clustering Analysis of Disease Prevalence Trends**

♦*Chenjin Ma1*.

1Beijing University of Technology.

In biomedical research, the analysis of disease prevalence is of critical importance. While most of the existing prevalence studies focus on individual diseases, there has been increasing effort that jointly examines the prevalence values and their trends of multiple diseases. Such joint analysis can provide valuable insights not shared by individual-disease analysis. A critical limitation of the existing analysis is that there is a lack of attention to existing information, which has been accumulated through a large number of studies and can be valuable especially when there are a large number of diseases but the number of prevalence values for a specific disease is limited. In this study, we conduct the functional clustering analysis of prevalence trends for a large number of diseases. A novel approach based on the penalized fusion technique is developed to incorporate information mined from published articles. It is innovatively designed to take into account that such information may not be fully relevant or correct. Another significant development is that statistical properties are rigorously established. Simulation is conducted and demonstrates its competitive performance. In the analysis of data from Taiwan NHIRD (National Health Insurance Research Database), new and interesting findings that differ from the existing ones are made.

**Session 24CHI120: Statistical Methods for Network Data**

**Network comparison by multivariate inference**

*Mingyu Qi1,* ♦*Tianxi Li2, Wen Zhou3*

1University of Virginia， 2University of Minnesota， 3Colorado State University Comparing networks of varying sizes is a pivotal statistical challenge, crucial for interpreting real-world data across fields such as biology, medicine, and social sciences. A robust comparison technique must be both statistically rigorous and produce results that are scientifically insightful. This presentation delves into comparing networks through multivariate inference of network motifs. We will illustrate that relying solely on single motif inference is insufficient for deriving meaningful conclusions; it’s imperative to incorporate both joint and conditional inference. We will adapt the subsampling technique to multivariate contexts, showcasing its potency in yielding more discernible test outcomes.

**Latent Network Structure Learning from High Dimensional Multivariate Point Processes**

♦*Biao Cai*1, *Jingfei Zhang*2, *Yongtao Guan*3

1City University of Hong Kong, 2Emory Universiy, 3Chinese University of Hong Kong

Learning the latent network structure from large scale multivariate point process data is an important task in a wide range of scientific and business applications. For instance, we might wish to estimate the neuronal functional connectivity network based on spiking times recorded from a collection of neurons. To characterize the complex processes underlying the observed data, we propose a new and flexible class of nonstationary Hawkes processes that allow both excitatory and inhibitory effects. We estimate the latent network structure using an efficient sparse least squares estimation approach. Using a thinning representation, we establish concentration inequalities for the first and second order statistics of the proposed Hawkes process. Such theoretical results enable us to establish the non-asymptotic error bound and the selection consistency of the estimated parameters. Furthermore, we describe a least squares loss based statistic for testing if the background intensity is constant in time. We demonstrate the efficacy of our proposed method through simulation studies and an application to a neuron spike train data set.

**Joint community detection in random effects stochastic block models via the split-likelihood method**

*Jiangzhou Wang*1, *Meier Wu*2, ♦*Binghui Liu*3, *Jianhua Guo*4

1Shenzhen University， 2Shenyang Agricultural University， 3Northeast Normal University， 4Beijing Technology and Business University

In this study, we tackle the joint community detection in multi-layer networks under a random effects stochastic block model. This model presents a unique challenge as it induces variability in the community structure across each layer of the multi-layer network. This variability is a random transformation originating from a common community structure that permeates all layers. The exact fit for this model is an NP-hard problem. We propose a solution, the ‘split-likelihood method’, which balances detection accuracy and computational efficiency. It employs an approximate likelihood maximization process by decoupling the row and column labels of community assignment. We further establish the convergence theory for our proposed method, along with the consistency theories for the estimated community labels derived from it. Extensive numerical results suggest that our method excels in both detection accuracy and computational efficiency. Finally, we conducted a resting state fMRI study on schizophrenia, to demonstrate the practical applicability of the proposed method.

**Adaptive Merging and Efficient Estimation in Longitudinal Networks**

*Haoran Zhang*1, ♦*Junhui Wang*2

1Southern University of Science and Technology， 2Chinese University of Hong Kong

Longitudinal network consists of a sequence of temporal edges among multiple nodes, where the temporal edges are observed in real time. It has become ubiquitous with the rise of online social platform and e-commerce, but largely under-investigated in literature. In this talk, we present an efficient estimation framework for longitudinal network, leveraging strengths of adaptive network merging, tensor decomposition and point process. It merges neighboring sparse networks so as to enlarge the number of observed edges and reduce estimation variance, whereas the estimation bias introduced by network merging is controlled by exploiting local temporal structures for adaptive network neighborhood. A projected gradient descent algorithm is proposed to facilitate estimation, where the upper bound of the estimation error in each iteration is established. Theoretical analysis of the proposed method shows that it can significantly reduce the estimation error and also provides guideline for network merging under various scenarios. We further demonstrate the advantage of the proposed method through extensive numerical experiments on synthetic datasets and a militarized interstate dispute dataset.

**Session 24CHI71: Recent Advances in Response Adaptive Randomization**

**The properties of covariate-adaptive randomization procedures with possibly unequal allocation ratio**

♦*Xiao Liu*1, *Feifang Hu*2 *and Wei Ma*1. 1Renmin University of China， 2George Washington University

In clinical trials, covariate-adaptive randomization (CAR) procedures are used to balance important covariates for more convincing results and enhanced statistical efficiency. Although most CAR procedures focus on trials with 1:1 allocation ratio, the demand for unequal allocation is growing. Therefore, this paper proposes a CAR framework that unifies numerous existing procedures and can balance general (discrete, continuous, or their combinations) covariates under any allocation ratio. To evaluate the proposed procedure, we classify covariates into randomized covariates and additional covariates, based on whether or not they are used in the randomization procedure. The analysis indicates that our proposed procedure possesses superior balancing properties for randomized covariates. Subsequently, we investigate the impact of CAR procedures on additional covariates. When balancing only discrete covariates, our results exhibit the benefit of CAR procedures in balancing additional covariates. However, the most intriguing finding is that, under unequal allocation ratio, balancing continuous covariates will challenge the balance of additional covariates, and we refer to this new issue as the `shift problem.' To understand and remedy this issue, we perform a comprehensive analysis about when and why it occurs, followed by two practical solutions to address the shift problem. The proposed CAR procedures are shown to effectively balance covariates when applied to the data from a depression trial.

**Bayesian adaptive randomization for heterogeneous causal effects**

♦*Zhongqiang Liu.*

Many adaptive randomization procedures have been proposed and studied over the past few decades. However, most of these procedures are based on frequentist constructions and can not apply to Bayesian situation. Here we introduce Bayesian-adaptive randomization technique, designed specifically for identifying heterogeneous treatment effects, while also considering essential statistical factors.

**Covariate adjustment for doubly adaptive biased coin design**

♦*Fuyi Tu*1 ， *Wei Ma*2.

1Chongqing University of Posts and Telecommunications， 2Renmin University of China

The doubly adaptive biased coin design has received great attention in both medical and biostatistical literature due to its advantages in efficiency improvement and ethical benefits. However, only patients’ responses are considered at the design stage, while substantial covariates are ignored. We propose a unified framework that effectively integrates covariate information by employing nonparametric and machine learning methods for treatment effect adjustment at the analysis stage. In addition, we present the asymptotic properties of the treatment effect estimator and provide consistent variance estimators for valid inference.

**Response-Adaptive Randomization Procedure in Clinical Trials with Surrogate Endpoints**

♦*Jingya Gao1*, *Feifang Hu2* and *Wei Ma3*

1University of Science， 2Technology Beijing， 3George Washington University, Renmin University of China

In clinical trials, subjects are usually recruited sequentially. Based on accumulated outcomes amassed thus far in the trial, the response-adaptive randomization (RAR) design has been shown to be a superb treatment assignment tool which is capable to skew the treatment allocation proportion according to pre-specified objectives, such as sending more patients to the more promising treatment. Unfortunately, there are circumstances that very few responses on the primary endpoints have been collected during recruitment period, such as responses related to public health emergencies or chronic diseases, and hence RAR is difficult to be applied to allocate treatments by using available outcomes. To overcome this problem, if an informative surrogate endpoint can be acquired much earlier than the primary endpoint, the surrogate endpoint can be used as a substitute of the primary endpoint in the RAR procedure. In this paper, we propose a RAR procedure relies only on surrogate endpoints. The validity of statistical inference and ethical advantages are justified by both theories and simulations. Further, different types of surrogate endpoints and primary endpoints are considered. The results support the use of RAR with surrogate endpoints in clinical trials when primary outcomes could not be collected in a relative short period of time.

**Session 24CHI113: Statistical Frontiers: Exploring Novel Techniques in Modeling, Testing and Beyond**

**Bayes-Optimal Fair Classification with Linear Disparity Constraints via Pre-, In-, and Post-processing**

♦*Xianli Zeng*1, *Guang Cheng*2 *and Edgar Dobriban*1.

1University of Pennsylvania， 2University of California, Los Angeles

Machine learning algorithms may have disparate impacts on protected groups. To address this, we develop methods for Bayes-optimal fair classification, aiming to minimize classification error subject to given group fairness constraints. We introduce the notion of linear disparity measures and bilinear disparity measures. We show that several popular disparity measures---the deviations from demographic parity, equality of opportunity, and predictive equality---are bilinear. We find the form of Bayes-optimal fair classifiers under a single linear disparity measure, by uncovering a connection with the Neyman-Pearson lemma. For bilinear disparity measures, Bayes-optimal fair classifiers become group-wise thresholding rules. Our approach can also handle multiple fairness constraints (such as equalized odds), and the common scenario when the protected attribute cannot be used at the prediction phase. Leveraging our theoretical results, we design methods that learn fair Bayes-optimal classifiers under bilinear disparity constraints. Our methods cover three popular approaches to fairness-aware classification, via pre-processing (Fair Up- and Down-Sampling), in-processing (Fair Cost-Sensitive Classification) and post-processing (a Fair Plug-In Rule). Our methods control disparity directly while achieving near-optimal fairness-accuracy tradeoffs. We show empirically that our methods compare favorably to existing algorithms.

**An Outlier Detection Test for the Functional Linear Model**

♦*Jia Guo.*

We consider the outlier detection problem for the linear model with functional responses and time-independent covariates. A new statistic based on the centered L2-norm of standardized jackknife residual functions is studied. For Gaussian data, we propose a test for outliers using the three-cumulant matched chi-square-approximation method. For non-Gaussian data, we propose a test for outliers using bootstrap. A graphical method for checking whether the data are Gaussian is also provided.

**A New Maximum Mean Discrepancy Based Two-sample Test for Equal Distributions in Separable Metric Spaces**

♦*Bu Zhou*1, *Zhi Peng Ong*2 *and Jin-Ting Zhang*2.

1Zhejiang Gongshang University，2National University of Singapore

We propose a novel two-sample test for equal distributions in separable metric spaces, utilizing the Maximum Mean Discrepancy (MMD). The test statistic is derived from the decomposition of the total variation of data in the reproducing kernel Hilbert space, and can be regarded as a V-statistic-based estimator of the squared MMD. We establish the asymptotic null and alternative distributions of the test statistic. Additionally, we introduce a new data-adaptive method based on the median absolute deviation to select the kernel width of the Gaussian kernel, and a new permutation test combining two different Gaussian kernel width selection methods. Fast implementation of the test using matrix calculation is discussed.

**A Variable Selection Tree and Its Random Forest**

*Yu Liu*1, *Xu Qin*1, ♦*Zhibo Ca*i2

1University of Electronic Science and Technology of China, 2Renmin University of China

The Sure Independence Screening (SIS) provides a fast and efficient ranking for the importance of variables for ultra-high dimensional regressions. However, classical SIS cannot eliminate false importance in the ranking, which is exacerbated in nonparametric settings. To address this problem, a novel screening approach is proposed by partitioning the sample into subsets sequentially and creating a tree-like structure of sub-samples called the SIS-tree. SIS-tree is straightforward to implement and can be integrated with various measures of dependence. Additionally, SIS-tree is extended to a forest with improved performance. Through simulations, the proposed methods are demonstrated to have great improvement comparing with existing SIS methods. The selection of a cutoff for the screening is also investigated through theoretical justification and experimental study. As a direct application of the screening, the classification of high-dimensional data is considered, and it is found that the ranking and cutoff can substantially improve the performance of existing classifiers.

**Session 24CHI10: Advances in Causal Inference: Methods and Applications**

**Fighting fire with fire: a synthetic control analysis of low-intensity burns' role in mitigating wildfire risk**

♦*Xiao Wu*1, *Erik Sverdrup*2, *Michael Mastrandrea*2, *Michael Wara*2 *and Stefan Wager*2.

1Columbia University， 2Stanford University

The increasing frequency of severe wildfires across the globe demands a shift in landscape management to mitigate their consequences. The role of managed, low-intensity fire as a driver of beneficial fuel treatment in fire-adapted ecosystems has drawn interest in both scientific and policy venues. Using a synthetic control approach to analyze twenty years of satellite-based fire activity data across 124,186 km2 of forests in California, we provide evidence that low-intensity fires substantially reduce the risk of future high-intensity fires. In conifer forests, the risk of high-intensity fire is reduced by 64.0% [95% CI: 41.2%–77.9%] in areas recently burned at low intensity relative to comparable unburned areas, and protective effects last for at least six years [lower bound of one-sided 95% CI: 6yr]. These findings support a transition from policies focused on fire suppression to ones emphasizing restoration, through increased use of prescribed fire, cultural burning, and managed wildfire, of a fire regime that approximates pre-suppression, pre-colonial conditions in California.

**Efficient estimation of causal effects when combining heterogeneous data with and without instrumental variables**

♦*Wei Li*1, *Peng Ding*2 *and Zhi Geng*3. 1Renmin University of China， 2University of California, Berkeley，3Beijing Technology and Business University

The increasing availability of multiple data sources has opened up new opportunities for causal inference in observational studies. This paper addresses estimation of causal effects in a primary population of interest with unmeasured confounders by leveraging auxiliary data with a valid instrumental variable. Existing data fusion methods often require homogeneity assumptions or, at the very least, the transportability of the conditional average treatment effect across the two populations. In this paper, we introduce a novel assumption termed unmeasured confounding exchangeability to ensure identification even when the two populations exhibit heterogeneous conditional average treatment effects. We then construct a multiply robust estimator that remains consistent in the presence of partial misspecifications of the observed data model, and it achieves local efficiency if all nuisance models are correct. We also develop a flexible sensitivity analysis framework for the assumption of unmeasured confounding exchangeability. The proposed approaches are illustrated through simulations and an application evaluating the causal effect of job satisfaction on smoking in China.

**Long-term causal inference under persistent confounding via data combination**

*Guido Imbens*1, *Nathan Kallus*2, ♦*Xiaojie Mao*3 *and Yuhao Wang*3.

1Stanford University， 2Cornell University， 3Tsinghua University

We study the identification and estimation of long-term treatment effects when both experimental and observational data are available. Since the long-term outcome is observed only after a long delay, it is not measured in the experimental data, but only recorded in the observational data. However, both types of data include observations of some short-term outcomes. In this paper, we uniquely tackle the challenge of persistent unmeasured confounders, i.e., some unmeasured confounders that can simultaneously affect the treatment, short-term outcomes and the long-term outcome, noting that they invalidate identification strategies in previous literature. To address this challenge, we exploit the sequential structure of multiple short-term outcomes, and develop three novel identification strategies for the average long-term treatment effect. We further propose three corresponding estimators and prove their asymptotic consistency and asymptotic normality. We finally apply our methods to estimate the effect of a job training program on long-term employment using semi-synthetic data. We numerically show that our proposals outperform existing methods that fail to handle persistent confounders.

**An exhaustive selection of sufficient adjustment sets for causal inference**

*Wei Luo*1*,* ♦*Fei Qin*2 *and Li-Xing Zhu*3*.*

1Zhejiang University， 2HongKong Baptist University， 3Beijing Normal University at Zhuhai

A subvector of predictor that satisfies the ignorability assumption, whose index set is called a sufficient adjustment set, is crucial for conducting reliable causal inference based on observational data. In this paper, we propose a general family of methods to detect all such sets for the first time in the literature, with no parametric assumptions on the outcome models and with flexible parametric and semiparametric assumptions on the predictor within the treatment groups; the latter induces desired sample-level accuracy. We show that the collection of sufficient adjustment sets can uniquely facilitate multiple types of studies in causal inference, including sharpening the estimation of average causal effect and recovering fundamental connections between the outcome and the treatment hidden in the dependence structure of the predictor. These findings are illustrated by simulation studies and a real data example at the end.

**Session 24CHI22: Challenges in Random Field-Based Modelling and Computations in Spatial statistics**

**Deep Learning for Spatial Data through Low-Rank Approximations**

♦*Hao Zhang1*.

1Michigan State University

Kriging, a geostatistical method for optimal linear prediction, is widely used in various domains including mining, hydrology, and environmental and health sciences, as well as in computational experiments. This technique bears a close resemblance to kernel methods in machine learning. In this talk, we provide an alternative and deep neural network-based approach for spatial prediction, designed to be resilient against variations in both the mean function and the kernel function. A key technique is to numerically approximate the eigenfunctions in a Karhunen-Loeve expansion of the spatial process.

**Estimation of expected Euler characteristic curves of nonstationary smooth random fields**

*Fabian Telschow1,* ♦*Dan Cheng2, Pratyush Pranav3* and *Armin Schwartzman4 .*

1Humboldt-Universität zu Berlin， 2Arizona State University， 3Centre de Recherche Astrophysique de Lyon， 4University of California San Diego

The expected Euler characteristic curve (EEC) summarizes the topology of the excursion sets of a random fieldabove the excursion threshold in terms of its expected Euler characteristic. For large thresholds, the EEC is anexcellent approximation for the tail distribution of the supremum of a smooth Gaussian field, and has applications inthe control of familywise error rate (FWER) and construction of simultaneous confidence bands . Therefore, it isimportant and valuable to estimate the EEC. Viewed as a function of the excursion threshold, the EEC of a Gaussian-related field is expressed by the Gaussian kinematic formula as a finite sum of known functions multiplied by theLipschitz–Killing curvatures (LKC) of the generating Gaussian field. This transforms the estimation of EEC toestimating LKC. In this talk, I will present a new method to estimate the LKC as linear projections of “pinned” Eulercharacteristic curves obtained from realizations of Gaussian fields. This provides an efficient and accurate tool toestimate the EEC and hence high excursion probabilities of Gaussian fields.

**Vector random fields over the arccos-quasi-quadratic metric space**

♦*Chunsheng Ma1*.

1Wichita State University

We introduce the arccosine-qusi-quadratic metric space on a subset of $R^{d+1}$, such as a sphere, a ball, a simplex, or an elliptical surface,

with the metric being the composition of arccosine and qusi-quadratic functions. This metric is not only conditionally negative definite but also a measure definite kernel, and metric space incorporate several important cases in a unified framework so that we are able to construct and study metric-dependent random fields on different metric spaces in a unified manner.

**Model-Free Feature Selection for Ultra-High Dimensional Data with Measurement Error**

♦*Weixing Song.*

In this talk, we explore a feature screening process designed for ultra-high dimensional data compromised by measurement error. This method is an extension of the existing cumulative divergence association measure, originally developed for error-free data. By correcting the bias introduced by measurement errors, our procedure retains all informative features for further statistical analysis with significantly high probabilities. Unlike existing screening methods that deal with measurement error, our bias-corrected screening approach is capable of processing a broader range of data, extending beyond those with linear structures. We establish sure independence screening and rank consistency under relatively lenient conditions. The effectiveness of our method is demonstrated through numerical studies and real-world data applications, showcasing its superior performance compared to existing methods.

**Session 24CHI23: Complex Data: Methodology and Application**

**Addressing the odds ratio inflation issues in logistic regression with perfect separation**

♦*Chenyu Liu1 , Xi Qiao1 , Ruitao Liu1* and *Liangliang Zha1* .

Logistic regression is a fundamental tool in statistical modeling and predictive analytics when analyzing categorical outcomes. Despite its versatility, logistic regression is susceptible to issues of perfect separation, which often result in inflated parameter estimates and unstable model predictions. Researchers have conventionally focused on addressing perfect separation in logistic regression post hoc, after fitting the model. In contrast, we advocate for a pre hoc approach, utilizing metrics to assess the extent to which predictors are separated by the binary outcome. As such, we introduce an overlap index to quantify the degree of intersection between the fitted densities of the variable of interest across the two groups. Extensive simulations have revealed that the issue of perfect separation arises when the overlap index falls below 10%. Current approaches, like the Firth penalty, aim to mitigate this problem by shrinking coefficients towards zero. However, these methods often fail to accurately recover the true values of the coefficient estimates. Therefore, we developed a new Bayesian method that enhances the detection and correction of problematic coefficients. This method was rigorously tested through simulations and validated with real data from the Endometrial Cancer Study. Our findings show that this approach not only more accurately identifies the variables causing separation but also adjusts their coefficients more closely to their true values.

**Bayesian Spatial Transcriptomic Deconvolution using Graph Laplacian Prior**

*Jiasen Zhang1,* ♦*Weihong Guo1, Liangliang Zhang1* and *Xi Qiao1*.

1Case Western Reserve University

Spatial transcriptomics (ST) is a cutting-edge technique in the field of genomics that allows re-

searchers to analyze the gene expression profile of tissues while preserving their spatial information. It has lots of very exciting applications in fields including developmental biology, cancer research, neurobiology, and immunology. Starting from a 2D gene expression ST data and cell type composition (reference basis) matrix, ST deconvolution solves the problem of finding the representation coefficient matrix called cell type composition matrix of the gene expression data under the basis. We use a hierarchical Bayesian framework that incorporates spatial smoothness modeled by graph Laplacian. By defining a graph Laplacian based on both spatial cartesian distance and histological intensity, we can ensure locations that are spatial close and also have similar histological intensity to have similar cell type decomposition. Numerical results on both simulated and real data show significantly better results than methods such as nonnegative matrix factorization (NMF) and CARD. Bayesian framework also allows uncertainty quantification.

**Enhancing Missing Data Imputation through Combined Bipartite Graph and Complete Directed Graph**

♦*Ziqi Chen1 .*

1East China Normal University

In the field of missing data imputation: identifying and leveraging the critical interdependencies among features to improve the precision of feature imputation. We propose a novel framework, known as the bipartite and complete directed graph neural network (BCGNN). Within BCGNN, observations and features are differentiated as two distinct node types. The values of observed features are converted into attributed edges, linking nodes representing features with those representing observations. The bipartite segment of our framework inductively creates embedding representations for nodes, efficiently utilizing the comprehensive information encapsulated in the attributed edges. In parallel, the complete directed graph segment adeptly outlines and communicates the complex interdependencies among features, ensuring a deep understanding of their connections. When compared to contemporary leading imputation methodologies, BCGNN consistently outperforms them, achieving a noteworthy average reduction of 15% in mean absolute error across various feature imputation tasks. Our extensive experimental investigation confirms that an in-depth grasp of the interdependence structure substantially enhances the model’s capability to generate precise feature embeddings. We also highlight the model’s superior performance in label prediction tasks involving missing data, and its formidable ability to generalize to novel, unseen data points.

**The likelihood ratio test for structural changes in factor models**

*Jushan Bai1,* ♦*Jiangtao Duan2* and *Xu Han3* .

1Columbia University 2Xidian University 3City University of Hong Kong

A factor model with a break in its factor loadings is observationally equivalent to a model without changes in the loadings but with a change in the variance of its factors. This approach effectively transforms a high-dimensional structural change problem into a low-dimensional problem. This paper considers the likelihood ratio (LR) test for a variance change in the estimated factors. The LR test implicitly explores a special feature of the estimated factors: the pre-break and post-break variances can be a singular matrix under the alternative hypothesis, making the LR test diverging faster and thus more powerful than Wald-type tests. The better power property of the LR test is also confirmed by simulations. We also consider mean changes and multiple breaks. We apply this procedure to the factor modeling of the US employment and study the structural change problem using monthly industry-level data.

**Session 24CHI84: Recent Development in Statistical Process Monitoring**

**Directional fault classification for correlated High-Dimensional data streams using hidden Markov models**

♦*Dongdong Xiang1* .

1East China Normal University

Modern manufacturing systems are often installed with sensor networks which generate high-dimensional data at high velocity. These data streams offer valuable information about the industrial system’s real-time performance. If a shift occurs in the manufacturing process, fault diagnosis based on the data streams becomes a fundamental task as it identifies the affected data streams and provides insights into the root cause. Existing fault diagnostic methods either ignore the correlation between different streams or fail to determine the shift directions. In this paper, we propose a directional fault classification procedure that incorporates the between-stream correlations. We suggest a three-state hidden Markov model that captures the correlation structure and enables inference about the shift direction. We show that our procedure is optimal in the sense that it minimizes the expected number of false discoveries while controlling the proportion of missed signals at a desired level. We also propose a deconvolution-expectation-maximization (DEM) algorithm for estimating the model parameters and establish the asymptotic optimality for the data-driven version of our procedure. Numerical comparisons with an existing approach and an application to a semiconductor production study show that the proposed procedure works well in practice.

**Dynamic Modeling and Online Monitoring of Tensor Data Streams with Application to Passenger Flow Surveillance**

*Yifan Li, Chunjie Wu,* ♦*Wendong Li, Fugee Tsung* and *Jianhua Guo .*

Passenger flow surveillance in urban transport systems has emerged as a major global issue for smart city management. Governments are taking proper measures to monitor passenger flow in order to maintain social stability and to prevent unexpected group events. It is critical to develop a passenger flow surveillance system that continuously monitors the passenger flow over time and triggers a signal as soon as the passenger flow begins to deteriorate so that timely government intervention can be implemented. In this paper, passenger flow surveillance is novelly formulated as dynamic modeling and online monitoring of tensor data streams. Existing tensor monitoring methods either rely heavily on the assumption that the tensor coefficients exhibit a low-rank structure or are inapplicable to general-order tensors. We propose a unified monitoring framework based on the tensor normal distribution to overcome these challenges. We begin by developing a tensor model selection procedure that ensures that the chosen tensor structure strikes a balance between model complexity and estimation accuracy. Then, we propose an online estimation procedure to dynamically estimate the tensor parameters, on which sequential change-detection procedures using the generalized likelihood ratio test are proposed. Extensive simulations and an analysis of real passenger flow data in Hong Kong demonstrate the efficacy of our approach.

**Multivariate vs. multi-stream EWMA control charts**

♦*Sven Knoth1.*

1Helmut Schmidt University Hamburg

Multivariate EWMA control charts were introduced in Lowry et al. in 1992 and became a popular and effective tool for monitoring multivariate data. However, multi-stream data are somehow related to the aforementioned framework. In both cases, correlation between the components respective streams is considered. However, whereas the multivariate EWMA charts deploys a distance (Mahalanobis) in the multivariate space, the multi-stream EWMA chart comprises a set of univariate control charts. In this talk, we discuss feasible calculation of the detection performance of multi-stream EWMA charts (not many results are available so far) and compare their detection behavior to the better investigated multivariate EWMA charts. Essentially, numerical methods are applied. Extensive Monte Carlo studies confirm their validity.

**Multivariate Nonparametric Control Charts Based on Projection Pursuit**

♦*Jun Li*

Multivariate nonparametric control charts are highly sought-after due to their flexibility to adapt to different distribution assumptions. However, many of the existing multivariate nonparametric control charts are only distribution-free for certain distribution families. Although those distribution families may contain different distributions, it is still difficult to verify whether the underlying multivariate distribution from a particular application belongs to those distribution families in practice. A few existing multivariate nonparametric control charts are fully nonparametric. As shown in the literature and in our simulation studies, most of them are not efficient in detecting location shifts. In this talk, we propose a new multivariate nonparametric control chart based on the idea of projection pursuit. The proposed control chart is fully

**Session 24CHI100: Robust Learning Inference for Data Science**

**Robust estimation of number of factors in high dimensional factor modeling via Spearman's rank correlation matrix**

♦*Zeng Li1*

*1*Southern University of Science and Technology

Determining the number of factors in high-dimensional factor modeling is essential but challenging, especially when the data are heavy-tailed. In this paper, we introduce a new estimator based on the spectral properties of Spearman’s rank correlation matrix under the high-dimensional setting, where both dimension and sample size tend to infinity proportionally. Our estimator is applicable for scenarios where either the common factors or idiosyncratic errors follow heavy-tailed distributions. We prove that the proposed estimator is consistent under mild conditions. Numerical experiments also demonstrate the superiority of our estimator compared to existing methods, especially for the heavy-tailed case.

**SIMPLE-RC: Group Network Inference with Non-Sharp Nulls and Weak Signals**

*Jianqing Fan*1, *Yingying Fan*2, ♦*Jinchi Lv*2, *Fan Yang*3

1Princeton University， 2University of Southern California， 3Tsinghua University

Large-scale network inference with uncertainty quantification has important applications in natural, social, and medical sciences. The recent work of Fan, Fan, Han, and Lv (2022) introduced a general framework of statistical inference on membership profiles in large networks (SIMPLE) for testing the sharp null hypothesis that a pair of given nodes share the same membership profile. In real applications, there are often groups of nodes under investigation that may share similar membership profiles in the presence of relatively weaker signals than the setting considered in SIMPLE. To address these practical challenges, in this paper, we propose a SIMPLE method with random coupling (SIMPLE-RC) for testing the non-sharp composite null hypothesis that a group of given nodes shares similar (not necessarily identical) membership profiles under weaker signals. Utilizing the idea of random coupling, we construct our test as the maximum of the SIMPLE tests for subsampled node pairs from the group. Such a technique significantly reduces the correlation among individual SIMPLE tests while largely maintaining the power, enabling delicate analysis of the asymptotic distributions of the SIMPLE-RC test. Our method and theory cover both the cases with and without node degree heterogeneity. These new theoretical developments are empowered by a second-order expansion of spiked eigenvectors under the $\ell\_\infty$-norm, built upon our work for random matrices with weak spikes. Our theoretical results and the practical advantages of the newly proposed method are demonstrated through simulations and real data examples.

**Covariate-shift Robust Adaptive Transfer Learning for High-Dimensional Regression**

*Zelin He, Ying Sun,* ♦*Jingyuan Liu* and *Runze Li ,* *Jinchi Lv.*

The main challenge that sets transfer learning apart from traditional supervised learning is the distribution shift, reflected as the shift between the source and target models and that between the marginal covariate distributions. High-dimensional data introduces unique challenges, such as covariate shifts in the covariate correlation structure and model shifts across individual features in the model. In this work, we tackle model shifts in the presence of covariate shifts in the high-dimensional regression setting. Furthermore, to learn transferable information which may vary across features, we propose an adaptive transfer learning method that can detect and aggregate the feature-wise transferable structures. Non-asymptotic bound is provided for the estimation error of the target model, showing the robustness of the proposed method to high-dimensional covariate shifts.

**Wasserstein proximal coordinate gradient algorithms**

♦*Xiaohui Chen1, Rentian Yao2* and *Yun Yang2*.

1University of Southern California， 2University of Illinois at Urbana-Champaign

This paper concerns composite (geodesically) convex optimization over multiple distributions. The objective functional under consideration is composed of a convex potential energy, defined on a product of Wasserstein spaces (the space of all distributions with a finite second moment), and a sum of convex self-interaction and internal energies associated with each distribution. To efficiently solve this problem, we introduce the Wasserstein Proximal Coordinate Gradient (WPCG) algorithm. Under a Quadratic Growth (QG) condition on the objective functional, a condition more relaxed than the typical strongly convex requirement, WPCG is proven to attain exponential convergence to the unique global optimum. Implications regarding the choice of step size and update schemes (parallel, sequential and random) are also discussed. In the absence of the QG condition, WPCG is still demonstrated to converge to the global optimal solution, albeit at a slower polynomial rate. The algorithm and theoretical framework are applied to two representative examples: approximation Bayesian computation using mean-field variational approximation, and the computation of equilibrium in multi-species systems with cross-interaction. Numerical results for both examples are consistent with our theoretical findings.

**Session 24CHI110: Statistical Challenges in Medical Device Clinical Trials**

**Statistical Considerations in Clinical Trial Design with Composite Endpoints - Applications in Medical Device Trials**

♦*Yu Shu1 .*

1Abbott Labs

In this presentation, we will review the development of composite endpoints in a number of applications in Medical Device clinical trial, as well as statistical methods typically used to analyze such composite endpoints. Pros and cons of traditional non-hierarchical composite endpoint vs. hierarchical composite endpoints in the means of trial design and statistical analysis.

**Informative censoring in survival analysis**

♦*Chul Ahn1* .

1Edwards Lifesciences

The Kaplan-Meier estimator may overestimate the survival function if censoring and death are positively correlated, and underestimate the survival function if they are negatively correlated. In this talk, I will introduce two existing methods to address the informative censoring issues proposed by Slud & Rubinstein (1983, Biometrika) and William Link (1989, JASA), and compare them.

**Real-world data use in medical device regulatory submission**

♦*Yanglu Zhao1* .

1Edwards Lifesciences

Real-world data (RWD) and real world evidence use is gaining popularity in medical device regulatory submission recently. Several regulatory agencies, including FDA, NMPA, PMDA, etc. has published guidelines on RWD and RWE in regulatory submission. The lecture will discuss the feature of RWD, review current guidance on RWD use for submission purposes and introduce several statistical methods used to analyze RWD.

**Valid scientific evidence and regulatory considerations for medical devices**

♦*Xiwen Liao*1, *Chen Yao*1

1Peking University First Hospital, Peking University Clinical Research Institute

Within the fast-evolving landscape of medical technology, the reliance on traditional randomized controlled trials (RCTs) for the regulatory approval of medical devices has become increasingly controversial. Given the inherent complexity and variability of medical devices, device trials often face methodological and statistical challenges that can substantially hinder the applicability and effectiveness of RCTs. A paradigm shift should be encouraged to embrace a wider variety of valid scientific evidence, including real-world data and real-world evidence, in regulatory decision-making, in alignment with the U.S. Food and Drug Administration’s adaptive regulatory guidelines. This approach seeks a balance between fostering innovation and ensuring medical device safety and effectiveness, proposing a flexible and inclusive regulatory framework that better accommodates the unique characteristics of medical devices.

**Session 24CHI103: Some Recent Advances in Survival Analysis**

**A Conditional Approach for Regression Analysis of Case K Interval-Censored Failure Time Data with Informative Censoring**

♦*Mingyue Du1* and *Xingqiu Zhao2 .*

1Jilin University，2The Hong Kong Polytechnic University

This paper discusses regression analysis of case K interval-censored failure time data, a general type of failure time data, with informative censoring with the focus on simultaneous variable selection and estimation. Although many authors have considered this challenging problem, most of the existing methods assume independent or non-informative censoring. More importantly, they are frailty model-based approaches and cannot directly assess the degree of informative censoring among other shortcomings. To address these, we propose a conditional approach and develop a penalized sieve maximum likelihood procedure for the simultaneous variable selection and estimation of covariate effects. Furthermore,we establish the oracle property of the proposed method and illustrate the appropriateness and usefulness of the approach using a simulation study. Finally we apply the proposed method to a set of real data on Alzheimer’s disease and provide some new insights.

**Semiparametric Analysis of Generalized Accelerated Hazards Models under Informatively Double-Censored Data**

*Rui Ma*1, ♦*Shuying Wang*2, *Jianguo Sun*3 and *Shishun Zhao*4.

1Northeast University, 2Changchun University of Technology, 3University of Missouri and 4Jilin University

Double-censored data arise from many clinical studies when failure event of interest can only be observed within a specific interval or cannot be directly observed. Methods developed for the data are mainly base on common survival model like proportional hazards model. As a generalised alternative model, generalized accelerated hazards models consider a time lag for producing effects for individuals and crossing hazards situation, which also includes the proportional hazards, accelerated failure time and traditional accelerated hazards models as special cases. However, almost literatures neither consider the situation that the censoring mechanism may be informative. In the paper, we investigate semiparametric regression analysis under double-censored data with informative censoring. We proposed a frailty-based joint model for failure time of interest and observation process. And we consider the inference procedure with an EM-based algorithm. The estimation of variance under informative censoring is provided with a profile method. As the special case of informative censoring, we also consider the inference procedure with independent double-censored data. For this, we proposed a sieve maximum likelihood approach and provide the estimation of covariance matrix by solving the inverse hessian matrix. The proposed estimators of regression parameters under both informative and independent censoring are shown to be consistent and asymptotically normal and an extensive simulation study suggests that the proposed method works well. In addition, the proposed methods have been employed to the real data set from

an AIDS Clinical Trial which motivating this study.

**Regression Analysis of Complier Causal Treatment Effects under the Case-Cohort Studies with Interval-Censored Data**

*Yuqing Ma*1, ♦*Peijie Wang*1， *Jianguo Sun*2.

1Jilin University， 2University of Missouri.

It has attracted a great deal of interest in randomized survival studies to assess the causal treatment effect under non-compliance with time-to event outcome. However, sometimes the covariates are expensive to obtain and disease rate is low, so it may cost a lot to collect covariates from the full cohort. To deal with this problem, we consider applying the case-cohort strategy to causal inference studies, which produce interval censored failure time, a situation for which there does not seem to exist an established approach. A sieve inverse probability weighting estimation procedure is proposed and the resulting estimators are shown to be consistent and asymptotically normal. A simulation study is conducted to evaluate the finite sample performance of the proposed approach and suggests that it works well in practice. It is applied to a breast cancer screening study.

**Data Integration and Subsampling Techniques in Distribution Estimation for Event Times with Missing Origins**

♦*Yi Xiong*

Time-to-event data with missing origins often arises when the occurrence of the event is silent. For example, records of wildfires can only be collected after a fire has been reported and thus the exact time when the fire starts is unknown. To tackle this issue, Xiong et al. (2021) proposed an approach that synthesizes auxiliary longitudinal measures to aid the inference on the unobserved time origins via the first-hitting-time model. In this work, we consider using alternative auxiliary data, which is collected prior to the occurrence of the event, to tackle the issue of missing time origins. Motivated by the example of estimating distribution of duration for wildfires, we propose to use the preceding records of lightning strikes to aid inference of the ignition time and start time of a fire. We first integrate the lightning strikes data with the fire data via Kernel smoothing and then provide a distribution estimator for a fire’s ignition time. By viewing a fire’s start time subject to be censored within the interval of the ignition time and the report time, we further adjust Turnbull estimator with interval-censored data to estimate the distribution of missing origin. Driven by the large volume of lightning strikes data, we also adapt the proposed estimation procedures to sub-samples of the lightning strikes data. The proposed approach potentially has many applications. This research is a joint work with Professor Joan Hu in Simon Fraser University, Canada.

**Limit Theorems for A Supercritical Two-Type Decomposable Branching Process in a Random Environment**

*Yanqing Wang*1, ♦*Dianni Wang*1, *Jinling Liu*1, *Quansheng Liu*1

1Zhongnan University of Economics and Law

Let $Zn=(Zn^{(1)},Zn^{(2)})$ be a two-type decomposable branching process in an independent and identically distributed random environment, where a type $1$ particle may produce particles of types 1 and 2, while a type 2 particle can only give birth to type 2 particles. We consider asymptotic properties of this process in the supercritical case. Because $Zn^{(1)}$ is an usual single-type branching process in a random environment, we only consider $Zn^{(2)}$. First, under some moment conditions, we find a suitable factor $\Pin$ such that the normalized population size $Wn=\frac{Zn^{(2)}}{\Pin}$

converges almost surely to a finite random variable $W$, and provide a decomposition expression and a non-degeneracy condition of $W$. Second, we give conditions under which $(Wn)$ is convergent in $L^p$ for $p\geq1$, and bounded in $L^p$ for $0<p<1$. Finally, we establish a central limit theorem for $\log Zn^{(2)}$.

**Session 24CHI20: Big Data and Nonparametric Time Series Theory and Applications**

**Additive nonparametric models with multivariate stationary and nonstationary variables and time trend**

*Chaohua Dong*1 and ♦*Xinqi Wu*1. 1Zhongnan University of Economics and Law

This paper devotes to investigating the additive nonparametric models with time trend and a few stationary and nonstationary variables in order to cater for the diverse nature of variables in economics, finance, and related fields. As is well known, these three kinds of variables are commonly encountered in empirical analyses; the additive nonparametric models are a compromise between parametric linear models and nonparametric models, and they possess an enormous advantage in eschewing the notorious ``curse of dimensionality'' in nonparametric literature. Therefore, studied along with a variety of variables the models have broad applicability.

All unknown nonparametric functions are estimated by the orthogonal series methods under identification conditions, and we jointly establish pointwise asymptotic distribution theory for all estimators. We find that the convergence rates of the estimators for unknown functions of different type variables can be distinguished. Monte Carlo simulations are conducted, and the results demonstrate that the estimation method proposed performs well in finite sample circumstance. Finally, using our model an empirical study in exchange rate forecasting is implemented, and we compare the results with the AR(1) model, the ADL model, as well as the random walk model. We find that the proposed model in this paper outperforms these candidate models in terms of the out-of-sample mean squared errors, and exhibits better predictive performance.

**Model averaging of nonparametric additive models with time trend based on series methods**

*Chaohua Dong*1 and ♦*Chen Zhou*1. 1Zhongnan University of Economics and Law

This paper considers additive nonparametric models with time trend and uses orthogonal series methods to estimate all unknown functions. Since there are many candidate models that satisfy the theoretical requirement for series estimator, we apply the model averaging technique by weighting all estimators from valid candidate models, where the weights are selected by minimizing integrated mean squared error (IMSE). The model averaging estimator we obtained are proved to be asymptotically optimal. The results of the Monte Carlo simulations confirm the theoretical results, and our method is superior to its competitors. We used our methodology to study the impact of China's geopolitical risks on China's agricultural futures prices.

**Forecasting Arctic Sea Ice Extent in the Long-Run: A Nonlinear Approach**

♦*Li Chen*1 , *Jiti Gao*2 and *Farshid Vahid*2.

1Xiamen University 2Monash University

As a major consequence of global warming, the extent of Arctic sea ice has been shrinking in the past several decades. We establish nonlinear econometric models to estimate the long-run relationship between sea ice extent, global and hemispheric mean temperatures, and total emissions of greenhouse gases. By doing so, we provide long-range projections of Arctic sea ice extent conditioning on future pathways of greenhouse gas emissions and therefore forecast the emergence of an ice-free Arctic.

**Threshold expectile regressions with an unknown threshold for dependent data**

♦*Feipeng Zhang*1 and *Yundong Tu*2.

1Xi'an Jiaotong University 2Peking University

This paper introduces a threshold expectile regression model with an unknown threshold for dependent data, which enables simple characterization of nonlinearity and heteroscedasticity in economic and financial applications. Profile estimation is proposed for the unknown parameters, and a sup-Wald test is developed to test the existence of the threshold effect at a fixed expectile level. Inference issues across multiple expectile levels are further considered, with a likelihood-ratio-type test designed to check for the presence of a common threshold value. Monte Carlo simulations demonstrate the nice finite sample performance of the proposed inference procedures. Finally, an empirical application demonstrates that the debt-to-GDP ratio has a heterogeneous threshold effect on the U.S. growth rate across the growth distribution.

**Session 24CHI15: Advancing Interpretability in Statistical and Machine Learning for Transcriptomics and Genomics**

**Towards interpretable models to predict DNA bendability**

*Brody Kendall1 and* ♦*Ji-Ping Wang1*

1Northwestern University

DNA bendability is a fundamental mechanical property that affects almost every function of DNA in biological processes. From DNA packing/storage to gene regulation, DNAs are required to bind to various proteins where the affinity of binding largely determined by intrinsic DNA bendability. Recently a high throughput assay named loop-seq has been developed to quantify the intrinsic bendability of a massive number of DNA fragments simultaneously. Using the most recent loop-seq data, we have developed a software tool, DNAcycP, based on a deep-learning approach for intrinsic DNA bendability prediction. In this talk we present follow-up research on more interpretable models for DNA bendability prediction. The results from the new models shed new insight into important sequence features associated with DNA bendability.

**Denoising protein expression in droplet-based single-cell data**

*Ouyang Zhu1* *and* ♦*Jun Li1*

1University of Notre Dame

In single-cell biology, profiling the abundance of surface proteins with techniques such as CITE-seq aids in cell type identification. However, measurements of protein abundance by droplet-based single-cell profiling technologies often suffer from significant contamination by technical noise, which obscures true biological differences. Current computational strategies for denoising face various limitations, including a reliance on empty droplets or null controls, which are often unavailable; insufficient effectiveness due to ignoring protein-protein interactions; and low computational efficiency. Here, we introduce a new probabilistic model that employs a variational autoencoder to achieve solutions in a highly computationally efficient manner. Our method does not rely on empty droplets, facilitates information sharing across proteins, and has demonstrated superior performance in denoising across diverse datasets.

**Constrained variable selection with applications in statistical genomics**

♦*Hui Jiang1*

*1*University of Michigan

In recent years, constrained variable selection has found applications across diverse fields, from portfolio optimization to climate modeling. In this presentation, we will explore its utilization in statistical genomics, including the analysis of microbiome data and the identification of gene pairs as potential cancer biomarkers. Additionally, we will discuss some efficient algorithms tailored for fitting these models.

**Integrating multi-omics data for sparse latent space detection**

*Alex White*1, *Chi Zhang*1, ♦*Sha Cao*1

1Indiana University

The emergence of comprehensive matched -omics datasets offers unprecedented opportunities to explore sample heterogeneity in ways not possible through the analysis of individual -omic layers alone. However, the integration of these datasets poses significant challenges, primarily due to the high dimensionality of -omics data, difficulties in harmonizing disparate -omic layers that operate on different scales, and the often neglected regulatory relationships among these layers. To overcome these obstacles, we present a novel analytical framework: a two-way sparse mixture model of high-dimensional-to-high-dimensional regression. This approach is designed to enable interpretable clustering by integrating matched -omics data, thus overcoming common limitations related to data dimensionality, cross-scale disparities, and hidden regulatory interactions, which are frequently overlooked in existing methodologies. Through both simulation studies and real-data analysis, our method demonstrates superior performance in clustering accuracy and variable selection, highlighting its potential as a powerful tool for -omics data integration.

**Session 24CHI97: Recent Studies on Machine Learning: from Theory to Application**

**A Gaussian Mixture Model for Multiple Instance Learning with Partially Subsampled Instances**

♦*Baichen Yu*1, *Xuetong Li*1, *Jing Zhou*2 *and Hansheng Wang*1.

1Peking University 2Renmin University of China

Multiple instance learning is a powerful machine learning technique, which is found useful when numerous instances can be naturally grouped into different bags. Accordingly, a bag-level label can be created for each bag according to whether the instances contained in the bag are all negative or not. Thereafter, how to train a statistical model with bag-level labels with/without partially labeled instances becomes the problem of great interest. To this end, we develop a Gaussian mixture model (GMM) framework to describe the stochastic behavior of the instance-level feature vectors. Both the instance-based maximum likelihood estimator (IMLE) and the bag-based maximum likelihood estimator (BMLE) are theoretically investigated. We found that the statistical efficiency of the IMLE could be much better than that of the BMLE, if the instance-level labels are relatively hard to be predicted. To fix the problem, we develop here a subsampling-based maximum likelihood estimation (SMLE) approach, where the instance-level labels are partially provided through carefully subsampling. This leads to a significantly reduced labeling cost with little sacrifice in terms of statistical efficiency. To demonstrate the finite sample performance, extensive simulation studies are presented. A real data example using whole-slide images (WSIs) to diagnose metastatic breast cancer is illustrated.

**CNN aided slit-lamp image analysis in corneal disease diagnosis**

♦*Yang Yu*1, *Xiao Wang*1, *Dongfang Li*2 *and Zhen Guo*2.

1Qingdao University， 2Qingdao Eye Hospital of Shandong FirstMedical University

The application of CNN to slit-lamp images for the diagnosis of corneal diseases represents a novel and innovative approach for diagnosing and assessing corneal diseases. Our model is among the first to demonstrate the potential of deep learning techniques for analyzing slit-lamp images and accurately diagnosing a diverse array of corneal diseases, including bullous keratopathy, corneal ulcer, conjunctival papilloma, keratoconus, lattice corneal dystrophy, peripheral corneal degeneration, band keratopathy, and pterygium. The novelty of our approach lies in the development of a single, unified model that can accurately differentiate between multiple corneal diseases based on slit-lamp images, thus addressing the challenges of overlapping clinical manifestations and complex corneal structures. By incorporating transfer learning and data augmentation techniques, our model demonstrates robust performance across different patient populations and imaging devices, enhancing its potential for real-world clinical applications.

**Revisiting ResNet: a convolutional sparse coding-based interpretation for residual architecture**

♦*Xiaohui Yang*1.

1Henan University

The residual architecture has been widely used in convolutional neural networks. However, the theoretical interpretation of the effectiveness of residual is not yet clear. Additionally, there are few researchers who have studied the problem of what kind of residual structure is more natural and reasonable. To address these issues, this work focuses on theoretical interpretability analysis of residual from the perspective of convolutional sparse encoding. Furthermore, a single-layer residual structure is constructed and designed as an implicit layer, which can be flexibly integrated in various mainstream networks, and achieve good representation learning ability and interpretability. Experimental results show that the ResNet, YOLO, U-Net, GAN, VAE, as well as lightweight networks such as MobilNet, with the proposed module all can achieve good results in corresponding downstream tasks, such as classification, object detection, segmentation, super-resolution reconstruction and multimodal interpolation.

**Scalable Kernel k-Means with Randomized Sketching: From Theory to Algorithm**

♦*Rong Yin.*

Kernel $k$-means is arguably one of the most common approaches to clustering. We investigate the efficiency of kernel $k$-means combined with randomized sketches in terms of both statistical analysis and computational requirements. More precisely, we propose a unified randomized sketches framework to kernel $k$-means and investigate its excess risk bounds, obtaining the state-of-the-art risk bound with only a fraction of computations. Indeed, we prove that it suffices to choose the sketch dimension $\Omega(\sqrt{n})$ to obtain the same accuracy of exact kernel $k$-means with greatly reducing the computational costs, for sub-Gaussian sketches, the randomized orthogonal system (ROS) sketches, and Nystr\"{o}m kernel $k$-means, where $n$ is the number of samples. To the best of our knowledge, this is the first result of this kind for unsupervised learning. Finally, the numerical experiments on simulated data and real-world datasets validate our theoretical analysis.

**Session 24CHI19: Application of Causal Inference Approaches in Global Drug Development**

**A Hybrid Prior Bayesian Method for Combing Real-world data and External trial data in Bridging Studies**

*Keer Chen1, Zengyue Zheng1, Rui Chen1, Pengfei Zhu2, Shuping Jiang2,,Nan Li2,Weiwei Zhao2,Jinmei Chen3 and* ♦*Ying Wu3.*

1Southern Medical University， 2MSD R&D (China)， 3Southern Medical University

**Bayesian Estimation on Continuous-Time Marginal Structural Models with Unmeasured Confounders**

*Haiyan Zhu and* ♦*Yingchun Zhou.*

Current causal inference frameworks in longitudinal studies typically assume discrete time steps for time advancement. However, the presence of continuous-time longitudinal data and unmeasured confounders in modern medical studies present challenges to these frameworks, potentially rendering them invalid or inefficient in utilizing available data. This paper delves into the application of Bayesian estimation on Marginal Structural Models in continuous-time settings with unmeasured confounders. The focus lies on scenarios where interventions change in continuous-time scale while unmeasured confounders remain static. To identify the crucial treatment timepoint and accurately determine causal parameters, the paper introduces Hawkes processes to construct the outcome model. The proposed approach is compared to naive and continuous-time marginal structural models using inverse probability of treatment (IPT) weighting on simulated data, and is also applied to analyze the relationship between oxytocin use for labor augmentation and the risk of Postpartum Hemorrhage.

**Causal estimands and inference for a principal stratum of adherers in clinical trials**

*Yongming Qu1 and* ♦*Hongying Li1.*

1Eli Lilly

This A hybrid Prior Bayesian Method for Combing Real-world data and External trial data in Bridging Studies.

**Practical considerations of applying propensity score in answering causal questions when leveraging external evidence**

♦*Yuan Tian1.*

1Novartis Institutes for Biomedical Research Co., Shanghai, China

The use of external evidence has become more and more common in drug development. The appropriate use of external data could benefit clinical trial in various dimensions, such as improve the efficiency of the study design, expediate the process of the drug development, etc. When it brings the benefits, the challenges of handling different sources of bias emerges when the trial is not protected by randomization. Causal inference methods have been used in such trials. Among these methods, propensity score is one of the well accepted and commonly used techniques. However, the best practice and communication on these methods are not fully explored. In this talk, we will explore the causal questions to answer when we are leveraging the external evidence, including clear articulation of assumption under which causal conclusions can be obtained that helps to assess validity of results, the practice of using propensity score methods to estimate causal effects based on these assumptions, and the communication on these methods.

**Session 24CHI27: Computational Criminal Law**

**Mixture Conditional Regression with Ultrahigh Dimensional Text Data for Estimating Extralegal Factor Effects**

♦*Jiaxin Shi1, Fang Wang2, Yuan Gao1, Xiaojun Song1 and Hansheng Wang1.*

1Guanghua School of Management, Peking University， 2Data Science Institute, Shandong University

Testing judicial impartiality is a problem of fundamental importance in empirical legal studies, for which standard regression methods have been popularly used to estimate the extralegal factor effects. However, those methods cannot handle control variables with ultrahigh dimensionality, such as those found in judgment documents recorded in text format. To solve this problem, we develop a novel mixture conditional regression (MCR) approach, assuming that the whole sample can be classified into a number of latent classes. Within each latent class, a standard linear regression model can be used to model the relationship between the response and a key feature vector, which is assumed to be of a fixed dimension. Meanwhile, ultrahigh dimensional control variables are then used to determine the latent class membership, where a na\"ive Bayes type model is used to describe the relationship. Hence, the dimension of control variables is allowed to be arbitrarily high. A novel expectation-maximization algorithm is developed for model estimation. Therefore, we are able to estimate the key parameters of interest as efficiently as if the true class membership were known in advance. Simulation studies are presented to demonstrate the proposed MCR method. A real dataset of Chinese burglary offenses is analyzed for illustration purposes.

**“This Crime is Not That Crime” —— Classification and Evaluation of Four Common Crimes**

♦*Ke Xu1, Hangyu Liu1, Fang Wang2 ， Hansheng Wang3.*

1University of International Business and Economics， 2Shandong University 3Peking University

As the basis of criminal penalty, criminal conviction, integral to the protection of fundamental rights and freedom of people, constitutes the basis and the core issue of criminal trials. Based on the data published on China Judgments Online, we proposed two types of classification models to apply the data of four common crimes from China Judgments Online and expounded their applications in identifying “abnormal cases”, defined as wrongly sentenced cases in this paper. The two types of classification models we proposed are a two-stage model and two deep learning models. To construct the two-stage model, we first used three keyword-extraction models to extract the keywords and vectorize all the keywords, then used five classification models to build the two-stage model. For the deep learning models, we applied two different deep neural network models in the data to build the classifier. We then applied these two types of classification models to discover “abnormal cases” in two steps. In the first step, we applied the two-stage model to extract the “important words” that will significantly improve the probability of the two-stage model to classify cases into crimes of intentional injury. In the second step, we constructed a validation data set of cases whose verdicts are changed in the second instance rulings to test the “important words” extracted in first step and the ability of the two-stage model and the two deep learning models to discover “abnormal cases”. The results of this exercise show that: 1) “important words” extracted in the first step are often associated with “abnormal cases”; 2) these two types of classification models can effectively discover “abnormal cases”, but compared with the two deep learning models, the two-stage model (aka. TF-IDF & ANN, the combination of a keyword extraction model and a classic machine-learning model) is more capable of discovering “abnormal cases”.

**Digital Intelligent Public Prosecutor: an Application of Knowledge Graph Models**

♦*Xin Liu.*

Taking advantages of techniques in big data era, massive judicial documents have been collected by prosecutors. Consequently, judicial intelligence becomes feasible in the sense that a connection between case factors and judgment results may be learnt, provided with professional and sufficient labeled data from the document. In this talk, the knowledge map construction of judgment documents and the prediction technology of punishment is studied, by analyzing 4761 criminal judgments of dangerous driving crimes in Shanghai district 2020 to 2022. Not only is the key information extracted efficiently from a large number of unstructured text information of judgment documents, but also predict the punishment result of three aspects, i.e., length of detention, fine amount and whether a criminal will be suspended and sentenced. In terms of map construction, this paper uses the rule-based method to extract 76 structured entities including sentencing elements and punishment results from unstructured texts, and creates semantic relationships accordingly, so that a knowledge network is constructed. The experimental results show that the knowledge graph model in this work can effectively obtain structured information from unstructured judgment documents, obtain the relationship between entities, and realize the construction of knowledge graph with great accuracy. Such a method may provide prosecutors with reference for sentencing similar cases and reasonable sentencing suggestions.

**Nonlinear Saturated Adaptive Estimation Method in Sentencing Data Analysis**

The development of legal artificial intelligence has strongly contributed to the fairness and efficiency of the law. The use of artificial intelligence, as a core task in the construction of Smart Court Project, has helped improve the quality and efficiency of trials. However, legal artificial intelligence involves fundamental rights such as personal freedom and property, making the reliability and interpretability of the calculation results crucial. In view of this, we introduce a new method of nonlinear saturated adaptive estimation and apply it to sentencing data analysis. Specifically, based on the Criminal Law and sentencing guidelines, we propose a nonlinear saturated mechanism model with sentencing interpretability, and establish a Two-Step Quasi-Newton (TSQN) algorithm to identify unknown parameters recursively without ideal statistical data assumptions including the independent and identically distributed (i.i.d.) condition. Moreover, we provide theoretical guarantee for the reliability of parameter estimation for both finite and infinite length of data, which is convenient to handle different size of data. Compared with the traditional linear regression model and the neural network model, the sentencing computation results given by our nonlinear recursive identification theory are more consistent with the basic principles and rules of sentencing. Our results accurately reflect the impact and changes in sentencing factors and show better predictive ability for different crimes. Additionally, based on the above nonlinear recursive identification theory, we further conduct some empirical studies to analyze the implementation effectiveness of specific judicial policies, such as admitting guilt and accepting punishments, justifiable self-defense, etc. It is also worth mentioning that for elements extraction tasks, we properly utilize large language models (LLM) to design prompt project, which is efficient and precise for data structured preprocessing and can achieve an accuracy of over 90% even for complex judgment elements.

**Session 24CHI30: Efficient Analysis in Statistics and Related Fields**

**Smoothed empirical likelihood for the difference of two quantiles with the paired sample**

*Pangpang Liu1 ，* ♦*Yichuan Zhao.* 1Purdue University.

In this paper, we propose a novel smoothed empirical likelihood method for the difference of quantiles with paired samples. While the empirical likelihood for the difference of two quantiles with independent samples has been studied, it is crucial to develop a statistical procedure that accounts for the dependence between paired samples. To this end, we propose two estimating equations for the difference of two quantiles and introduce a nuisance parameter in our smoothed empirical likelihood framework. We demonstrate that our approach yields a limiting distribution that follows the standard \ch^2 distribution. Extensive simulation studies confirm that our smoothed empirical likelihood method outperforms the normal approximation and method M (Wilcox and Erceg-Hurn in J Appl Stat 39(12):2655–2664, 2012) in most cases. Finally, we illustrate the usefulness of our proposed method by applying it to a real-world data set, estimating the interval of the quantile difference of GDP between different years.

**Efficient estimation in homogeneous partially linear models**

*Wei Lin1 ，* ♦*Zhijian Li2.*

1Ohio University， 2Beijing Normal University-Hong Kong Baptist University United International College

The partially linear models (PLM) have been extensively examined in the literature when the errors are heteroscedastic. When the errors are homogeneous, on the other hand, most of the existing analysis assumes the independence between the error and the covariate vector. In this talk, we introduce the efficient estimator of the parameters in a homogeneous PLM without the independence aforementioned.

The asymptotic normality of a simpler semi-efficient error variance estimator is also established. An extended simulation study is conducted which demonstrates that our proposal performs very favorably against other estimators, especially when the error variance is relatively small.

**Sequential quantile regression for stream data by least squares**

*Ye Fan1 ，* ♦*Nan Lin2.*

1Capital University of Economics and Business， 2Washington University in St. Louis

Due to storage limitation, massive stream data cannot be permanently stored, and real-time analysis is often needed. We developed a sequential algorithm, SQR, to support efficient quantile regression (QR) analysis for stream data. Due to the non-smoothness of the check loss, popular gradient-based methods do not directly apply to QR. Motivated by the Bayesian QR, our proposed SQR algorithm is able to convert the non-smooth optimization into a least squares problem. Hence it is significantly faster than existing algorithms that all require solving a linear programming problem in local processing. We prove that the SQR estimator is unbiased, asymptotically normal and enjoys a linear convergence rate under mild conditions.

**Gene Classification with an OCCA-Based Unsupervised Feature Selection Framework**

♦*Wanjun Ning1.*

1University of Texas at Arlington

Processing high-dimensional data efficiently poses a significant challenge in today's big data era, particularly in the field of bioinformatics. Understanding the development of tumors in oncology research hinges on the identification of disease-related genes. With the increasing prevalence of high-dimensional genetic data, feature selection has gained paramount importance in both bioinformatics and medicine. Effective feature selection techniques are indispensable for unraveling the intricacies of tumor genomics.

To enhance the identification of crucial genes in tumors, our research introduces an unsupervised feature selection framework grounded in Orthogonal Canonical Correlation Analysis (OCCA). Unlike conventional methods that often focus solely on local or global data structures, OCCA allows us to simultaneously capture both while preserving the integrity of the data. This results in more precise and relevant gene selection for essential traits and facilitates more efficient analysis of large-scale bioinformatic data.

Our framework, as demonstrated in experiments with tumor datasets, outperforms existing unsupervised feature selection algorithms in pinpointing critical genes. This underscores the significance of amalgamating local and global data structures, providing fresh perspectives and innovative approaches to the study and treatment of tumor-related genes. This research holds particular relevance at the intersection of data science and biomedicine, especially in the context of advancing precision medicine and personalized therapy.

**Session 24CHI39: Innovations in Statistical Methodologies: From Functional Data Analysis to Spatial Modeling**

**Distributed Heterogeneity Learning for Generalized Partially Linear Models with Spatially Varying Coefficients**

♦*Shan Yu1, Guannan Wang2 ， Li Wang3.*

1University of Virginia， 2College of William and Marry， 3George Mason University

Spatial heterogeneity is of great importance in social, economic, and environmental science studies. The spatially varying coefficient model is a popular and effective spatial regression technique to address spatial heterogeneity. However, accounting for heterogeneity comes at the cost of reducing model parsimony. To balance flexibility and parsimony, this article develops a class of generalized partially linear spatially varying coefficient models which allow the inclusion of both constant and spatially varying effects of covariates. Another significant challenge in many applications comes from the enormous size of the spatial datasets collected from modern technologies. To tackle this challenge, we design a novel distributed heterogeneity learning (DHL) method based on bivariate spline smoothing over a triangulation of the domain. The proposed DHL algorithm has a simple, scalable, and communication-efficient implementation scheme that can almost achieve linear speedup. In addition, this article provides rigorous theoretical support for the DHL framework. We prove that the DHL constant coefficient estimators are asymptotic normal and the DHL spline estimators reach the same convergence rate as the global spline estimators obtained using the entire dataset. The proposed DHL method is evaluated through extensive simulation studies and analyses of the U.S. loan application data.

**Shape Mediation Analysis in Alzheimer’s Disease Studies**

*Xingcai Zhou, Miyeon Yeon,* ♦*Jiangyan Wang, Shengxian Ding, Kaizhou Lei, Rongjie Liu and Chao Huang.*

As a crucial tool in neuroscience, mediation analysis has been developed and widely adopted to elucidate the role of intermediary variables derived from neuroimaging data. Typically, structural equation models (SEMs) are employed to investigate the influences of exposures on outcomes, with model coefficients being interpreted as causal effects. While existing SEMs have proven to be effective tools for mediation analysis involving various neuroimaging-related mediators, limited research has explored scenarios where these mediators are derived from the shape space. In addition, the linear relationship assumption adopted in existing SEMs may lead to substantial efficiency losses and decreased predictive accuracy in real-world applications. To address these challenges, we introduce a novel framework for shape mediation analysis, designed to explore the causal relationships between genetic exposures and clinical outcomes, whether mediated or unmediated by shape-related factors while accounting for potential confounding variables. Within our framework, we apply the square-root velocity function to extract elastic shape representations, which reside within the linear Hilbert space of square-integrable functions. Subsequently, we introduce a two-layer shape regression model to characterize the relationships among neurocognitive outcomes, elastic shape mediators, genetic exposures, and clinical confounders. Both estimation and inference procedures are established for unknown parameters along with the corresponding causal estimands. The asymptotic properties of estimated quantities are investigated as well. Both simulated studies and real-data analyses demonstrate the superior performance of our proposed method in terms of estimation accuracy and robustness when compared to existing approaches for estimating causal estimands.

**Simultaneous inference for non-stationary functional time series**

*Leheng Cai1 and* ♦*Lijian Yang1.*

1Tsinghua University

This paper develops simultaneous inference tools for the time-varying mean function of non-stationary functional time series. To achieve this goal, An ``infeasible'' B-spline estimator is proposed with all random trajectories completely recorded without any errors, and a Gaussian approximation is derived for the standardized maximum deviation, leading to an ``infeasible'' simultaneous confidence region (SCR) of the time-varying bivariate mean function. A two-step data-driven estimator is proposed, equivalent to a computationally efficient tensor-product bivariate spline estimator, when trajectories are observed only on discrete points with measurement errors. Under mild conditions, this two-step estimator is oracally efficient in the sense that it enjoys the same asymptotic properties as the infeasible estimator, which allows one to construct an asymptotically correct SCR for the bivariate mean function. The SCRs do not have explicit form, but all enjoy an adaptive and uniform width that is a factor $\log^{1/2} T$ wider than pointwise confidence intervals from nonparametric regression. Various extensions are also studied, including SCR for marginal univariate mean functions, and we further test of additivity for the bivariate mean. Extensive simulation results strongly support the theoretical results, and the SCRs are applied to a fertility rate data and a temperature curve data.

**Partially functional linear quantile regression with measurement error**

*Mengli Zhang*1, ♦*Lan Xue*2, *Carmen Tekwe*3, *Yang Bai1* and Annie Qu4.

1Shanghai University of Economics and Finance, 2Oregon State University, 3Indiana University, 4University of California, Irvine.

Ignoring measurement errors in conventional regression analyses can lead to biased estimation and inference results. Reducing such bias could be challenging when the error-prone covariate is a functional curve. In this paper, we propose a new corrected loss function for a partially functional linear quantile model with function-valued measurement errors. We establish asymptotic properties of both functional-coefficient and parametric coefficient estimators. We also demonstrate the finite sample performance of the proposed method through simulation studies and illustrate its advantages for children obesity study.

**Session 24CHI40: Innovative Machine Learning Approaches in Genomics and Genetics**

**MAAT: a new nonparametric Bayesian framework for incorporating multiple functional annotations in transcriptome-wide association studies**

♦*Han Wang1, Teng Li2, Pak Chung Sham3 and Yan Dora Zhang3.*

1China Agricultural University， 2Peking Union Medical College， 3The University of Hong Kong

Transcriptome-wide association studies (TWAS) has emerged as a powerful tool for translating the myriad variations identified by genome-wide association studies (GWAS) into regulated genes in the post-GWAS era. While integrating annotation information has been shown to enhance power, current annotation-assisted TWAS tools predominantly focus on epigenomic annotations. When including more annotations, the assumption of a positive correlation between annotation scores and SNPs' effect sizes, as adopted by current methods, often falls short. Here, we propose MAAT (multiple annotation-assisted TWAS), expanding the horizons of existing TWAS studies in two pivotal ways: (i) We propose a non-parametric PPMx prior to incorporate information from seven multifaceted annotations into TWAS, getting free from the reliance on the assumption of a linear relationship between annotations and effect sizes. The included annotation also extends beyond epigenetic data. (ii) Beyond the scope of existing TWAS, we introduce an angle-based metric indicating which annotation plays the most crucial role when a gene influences a trait, providing new perspectives for understanding the biological mechanisms. Through simulations, we demonstrate MAAT outperforms existing state-of-the-art TWAS methods in terms of imputation R2 and association power. Applying MAAT to eight psychiatric traits, we identify more gene-trait associations and provide both validation and interpretation of the assigned annotations.

**Investigating spatial dynamics in spatial omics data with StarTrail**

*Jiawen Chen1, Caiwei Xiong1, Quan Sun1, Gaorav Gupta1, Aritra Halder2,* ♦*Yun Li1 and Didong Li1.*

1University of North Carolina, Chapel Hill， 2Drexel University

Spatial omics technologies revolutionize our view of biological processes within tissues. However, existing methods fail to capture localized, sharp changes characteristic of critical events (e.g. tumor development). Here, we present StarTrail, a novel gradient based method that powerfully defines rapidly changing regions and detects “cliff genes”, genes exhibiting drastic expression changes at highly localized or disjoint boundaries. StarTrail, the first to leverage spatial gradients for spatial omics data, also quantifies directional dynamics. Across multiple datasets, StarTrail accurately delineates cellular boundaries (e.g., brain layers, tumor-immune boundaries), and detects cliff genes that may regulate molecular crosstalk at these biologically relevant boundaries but are missed by existing methods. For instance, StarTrail precisely pinpointed the cancer immune interface in a HER2+ breast cancer dataset, unveiled key cliff genes including a potential prognostic biomarker, highlighting NK-, B-cell mediated immunity, and B cell receptor signaling pathways missed by all spatial variable gene methods attempted. StarTrail, filling important gaps in current literature, enables deeper insights into tissue spatial architecture.

**Seq2Karyotype (S2K): A Method for Deconvoluting Heterogeneity of Copy Number Alterations Using Single-Sample Whole-Genome Sequencing Data**

*Limeng Pu1,* ♦*Xiang Chen1 and Jinghui Zhang1.*

1St. Jude Children's Research Hospital

Broad copy number alterations (CNAs) at chromosomal band resolution have been used as cancer diagnostic and prognostic biomarkers for decades. These events have been characterized by cytogenetic imaging, an approach which is powerful for assessing heterogeneity but limited for locus fine mapping. While CNA detection by next-generation sequencing has become a standard analysis, existing methods rarely model CNA heterogeneity in bulk tumor samples and often require a paired normal sample for control. To overcome these limitations, we developed Seq2Karyotype (S2K, https://github.com/chenlab-sj/Seq2Karyotype), a new algorithm for in-silico karyotyping using single-sample whole-genome sequencing (WGS) data. S2K performs joint modeling of read-depth and allelic imbalance (AI) of high-quality heterozygous SNPs to identify reference diploid regions, followed by modeling and segmenting the deviation from the reference. Empirical coverage and AI of segmented regions are fitted to models of single and admixed CNAs to estimate clonality. The final karyotyping considers both the model fitness and minimization of evolution steps. To demonstrate S2K’s utility, we analyzed diverse data sets including both primary patient samples and tumor cell lines. The analysis results not only demonstrate the accuracy of in-silico karyotyping performed by S2K but also reveal the dynamic intra-tumor heterogeneity in cancer cell lines, which may impact the design and interpretation of future experiments using these cell lines.

**PURE: An Integrated Approach for Causal Protein Discovery Leveraging Cis- and Trans-acting Elements**

*Zichen Zhang*1, *Bingxin Zhao*2 *and* ♦*Chong Wu*1.

1UT MD Anderson Cancer Center， 2University of Pennsylvania

quantitative trait loci (pQTL) datasets integrated with genome-wide association studies (GWAS), are

frequently used to identify probable causal proteins. Analogous to transcriptome-wide association studies

(TWAS), current PWAS methods predominantly focus on cis-acting elements, constructing protein levels prediction models using individual-level pQTL datasets, often limited by sample size. However, trans-acting elements can account for a significant proportion of variation in many protein markers and often play essential regulatory roles. To maximize the potential of trans-acting elements and summary-level pQTL data for improving the robustness and power of PWAS, we introduce a novel PWAS method, referred to as Protein-trait association Using cis- and tRans-regulation Estimation (PURE). Our method involves two steps. First, we develop prediction models using each cis- and trans-acting element leveraging summarylevel data with large sample size, thereby addressing the relatively weak effects of trans-acting elements. Next, we estimate the associations between phenotype and genetically predicted protein levels for each cis- and trans-acting element. These associations, estimated from each locus, are then combined using an iterative algorithm to account for certain outliers and randomness in outlier detection. Applying our novel method to deCODE summary-level data, encompassing plasma protein levels of 4,907 proteins derived from 35,559 Icelanders, we constructed 2,127 protein prediction models demonstrating satisfactory performance (R2>0.01), achieving a 58% improvement over the existing model constructed using the ARIC dataset. Further external validation of our models using the INTERVAL data yielded a high validation rate. Finally, in a case study for Alzheimer’s disease using recent GWAS data of 111,326 clinically diagnosed/proxy cases and 677,663 controls, our model identified 207 likely causal proteins under Bonferroni correction. In contrast, competing methods identified 19 likely causal proteins. We will release a

companion software on GitHub to enable wider use of our method

**Session 24CHI42: Machine Learning and Causal Inference Advances in Biostatistics**

**CONTRASTIVE EIGEN ANALYSIS FOR EXPLORING HIDDEN EFFECTS**

♦*Yunhui Qi*1, *Yumou Qiu*2, *Peng Liu*1

1Iowa State University, 2Peking University

Contrastive principal component analysis (cPCA) and its sparse version (scPCA) have proven effective in learning the latent structure of datasets with multiple sample groups, such as treatment and control. While cPCA effectively identifies treatment-enriched latent subgroups, the theoretical underpinnings of subgroup detection remain unclear. The need for a deeper comprehension of variation trade-off tuning parameter determination and the development of automated selection techniques is also emergent. More importantly, few statistical methods have been developed to contrastively analyze multi-modal datasets. This work introduces a comprehensive contrastive eigen analysis framework, employing singular value decomposition and penalized matrix decomposition, within which, cPCA is theoretically analyzed to unveil the conditions for subgroup identification and nuisance effects removal. We theoretically derive the optimal tuning parameter values for subgroup identification and propose an ad-hoc method for hyper-parameter tuning without prior information. Furthermore, we introduce two innovative approaches, contrastive cross covariance analysis (3CA) and its sparse version (s3CA), for examining shifts in cross-modal data relationships from control to treatment groups. Theoretical analysis proves the adaptability of these methods in managing various error covariances and distinguishing between shared and unique factors. Demonstrated via simulation studies and applied to real-world datasets like barley microarray and COVID-19 multi-omics, our methods uncover significant insights into dynamic effect of genotype/pathogen and unique pathways associated with platelet counts in COVID-19 patients, showcasing their utility in advanced contrastive analysis.

**Biclustering methods with feature selection for omics data analysis**

♦*Peng Liu*1, *Zhili Qiao*1

1Iowa State University

With the increasing availability of sequencing datasets, the interest in discovering functionally related omics features under various experimental conditions has emerged. Bicluster analysis is an unsupervised data mining technique to meet this purpose by simultaneously clustering rows and columns of a data matrix. Omics data typically contain tens of thousands of features, with the majority of features most likely being non-informative. Selecting only informative features while doing cluster analysis can provide insights into the functions of features and improve biological interpretability. Although feature selection has been studied in one-dimensional cluster analysis, none has been done for biclustering. In this talk, we investigate feature selection methods in bicluster analysis by applying and extending two different classes of penalization methods. By introducing sparsity into the feature space, we simultaneously identify feature and sample groups and select informative features. We compare these methods through a variety of simulation studies and apply them to biological datasets to examine their clustering accuracy and computational efficiency and find that one method consistently performs better than the others.

**Random-effects split conformal interval prediction for machine learning algorithms**

*Danyang Zhang*1, ♦*Chong Wang*1

1Iowa State University

Machine-learning algorithms have seen widespread application across various domains, aiming to optimize performance metrics through training experiences. However, conventional algorithms often yield point predictions devoid of uncertainty measures, posing limitations in certain contexts. The conformal prediction framework offers a solution, furnishing reliable predictions by establishing valid prediction intervals for regression problems. Split conformal (SC) prediction emerges as a computationally efficient alternative, segregating modeling and calibration tasks to enhance efficiency. However, SC methods assume independent and identically distributed (i.i.d.) data; violations of this assumption, especially in hierarchical data structures like swine production, warrant tailored approaches. Thus, we propose a hierarchical prediction interval method within the SC framework, leveraging the mixed-effects machine learning model's capability to accommodate hierarchical data structures. Our process, exemplified with support vector regression (SVR), aims to narrow prediction intervals by incorporating subject-specific information. Furthermore, we explore the application of SC, CDFs-pooling, and single-subsampling approaches with SVR for comparison. This study fills a gap in mixed-effects machine learning and conformal prediction literature, demonstrating the efficacy of the proposed method through empirical and simulation studies.

**Quantifying the global mediation effect for nonsparse high-dimensional genomics mediators**

*Tianzhong Yang*1, *Zhiyu Kang*1, ♦*Chunlin Li*2

1University of Minnesota, 2Iowa State University

While many existing epidemiological studies have examined associations between alcohol and cardiovascular outcomes, less has been done to explore causal biological pathways and mechanisms of the observed associations at the molecular level. To investigate this relationship, we propose a new causal measure to quantify the mediating role of molecular phenotypes, such as DNA methylation, in bridging alcohol intake and cardiovascular outcomes. The challenge of estimating this measure is two-fold. First, since alcohol consumption is associated with genome-wide changes at the molecular level, it is biologically plausible that many omics mediators with weak but collectively considerable effects are involved in the pathway; however, existing methods are plagued by inconsistency in the presence of non-sparse mediators. To address this issue, we develop a method to estimate the proposed measure in such situations consistently. Second, many epidemiological studies use case-control sampling, which introduces ascertainment bias in mediation analysis. To correct this bias, we propose a method of moment motivated by heritability estimation. Finally, a significant challenge in this research is the potential for residual confounding in observational studies, which can seriously compromise the validity of scientific findings. We will briefly discuss the approach to correct the confounding bias.

**Session 24CHI46: Modern Statistical Methods for Time Series Data**

**No-lose Converging Kernel Estimation of Long-run Variance**

*Xu Liu*1 *and* ♦*Kin Wai Chan*1*.*

1The Chinese University of Hong Kong

Kernel estimators have been popular for decades in long-run variance estimation. To minimize the loss of efficiency measured by the mean-squared error in important aspects of kernel estimation, we propose a novel class of converging kernel estimators that have the “no-lose” properties including: (1) no efficiency loss from estimating the bandwidth as the optimal choice is universal; (2) no efficiency loss from ensuring positive-definiteness using a principle-driven aggregation technique; and (3) no efficiency loss asymptotically from potentially misspecified prewhitening models and transformations of the time series. A shrinkage prewhitening transformation is proposed for more robust finite-sample performance. The estimator has a positive bias that diminishes with the sample size so that it is more conservative compared with the typically negatively biased classical estimators. The proposal improves upon standard kernel functions and can be well generalized to the multivariate case. We discuss its performance through simulation results and a real-data application in MCMC convergence diagnostics.

**Asymptotics of sample tail autocorrelations for tail dependent time series: phase transition and** visualization

♦*Ting Zhang*1

*1*University of Georgia

We develop an asymptotic theory on sample tail autocorrelations of time series data that can exhibit serial dependence in both tail and non-tail regions. Unlike the traditional autocorrelation function, the study of tail autocorrelations requires a double asymptotic scheme to capture the tail phenomena, and our results do not impose any restriction on the dependence structure in non-tail regions and allow processes that are not necessarily strong mixing. Our asymptotic theory indicates a phase transition phenomenon for sample tail autocorrelations, whose asymptotic behavior including the convergence rate can transit from one phase to the other when the lag index moves past the point beyond which serial tail dependence vanishes. The phase transition fills the gap of existing research on tail autocorrelations, and can be used to construct the lines of significance, in analogy to the traditional autocorrelation plot, when visualizing sample tail autocorrelations to assess the existence of serial tail dependence or to identify the maximal lag of tail dependence.

**A Random Graph-based Autoregressive Model for Networked Time Series**

♦*Weichi Wu**and Chenlei Leng*

Contemporary time series data often feature objects connected by a social network that naturally induces temporal dependence involving connected neighbours. The network vector autoregressive model is useful for describing the influence of linked neighbours, while its recent generalizations aim to separate influence and homophily. Existing approaches, however, require either correct specification of a time series model or accurate estimation of a network model or both, and rely exclusively on least-squares for parameter estimation. This paper proposes a new autoregressive model incorporating a flexible form for latent variables used to depict homophily. We develop a first-order differencing method for the estimation of influence requiring only the influence part of the model to be correctly specified. When the homophily part is correctly specified admitting a semiparametric form, we leverage and generalize the recent notion of neighbour smoothing for parameter estimation, bypassing the need to specify the generative mechanism of the network. We develop new theory to show that all the estimated parameters are consistent and asymptotically normal. The efficacy of our approach is confirmed via extensive simulations and an analysis of a social media dataset.

**Statistical Inference For Functional Time Series**

♦*Jie Li*1 and *Lijian Yang*2.

1Renmin University of China， 2Tsinghua University.

We investigate statistical inference for the mean function of stationary functional time series data with an infinite moving average structure. We propose a B-spline estimation for the temporally ordered trajectories of the functional mov- ing average, which are used to construct a two-step estimator of the mean function. Under mild conditions, the B-spline mean estimator enjoys oracle efficiency in the sense that it is asymptotically equivalent to the infeasible estimator, that is, the sample mean of all trajectories observed entirely without errors. This oracle effi- ciency allows us to construct a simultaneous confidence band (SCB) for the mean function, which is asymptotically correct. Simulation results strongly corroborate the asymptotic theory. Using the SCB to analyze an electroencephalogram time series reveals strong evidence of a trigonometric form of the mean function.

**Detecting long-range dependence for time-varying linear models**

♦*Lujia Bai*

We consider the problem of testing for long-range dependence in time-varying coefficient regression models, where the covariates and errors are locally stationary, allowing complex temporal dynamics and heteroscedasticity. We develop KPSS, R/S, V/S, and K/S-type statistics based on the nonparametric residuals. Under the null hypothesis, the local alternatives as well as the fixed alternatives, we derive the limiting distributions of the test statistics. As the four types of test statistics could degenerate when the time-varying mean, variance, long-run variance of errors, covariates, and the intercept lie in certain hyperplanes, we show the bootstrap-assisted tests are consistent under both degenerate and non-degenerate scenarios. In particular, in the presence of covariates the exact local asymptotic power of the bootstrap-assisted tests can enjoy the same order as that of the classical KPSS test of long memory for strictly stationary series. The asymptotic theory is built on a new Gaussian approximation technique for locally stationary long-memory processes with short-memory covariates, which is of independent interest. The effectiveness of our tests is demonstrated by extensive simulation studies and real data analysis.

**Session 24CHI85: Recent Development in Survival Data and Event History Data Analysis**

**Accelerating the Elastic Net Penalized Cox Proportional Hazards Regression Using Safe Screening**

♦*Hong Wang1.*

1Central South University

The elastic net penalty which includes both the L1 and L2 penalties is a widely applied in survival analysis for simultaneous estimation and feature selection. Although many efforts have been devoted to its efficient implementation, its application to high dimensional or large sample sized survival data still poses significant computational challenges. In this paper, we present fast and effective safe screening rules for Cox proportional hazard model with (adaptive) elastic net penalties to identify the zero coefficients in the solution vector, which may lead to a substantial reduction in the number of features before the final optimization procedure. Theoretically, we prove that the algorithm with our screening rules are able to guarantee identical results with the original algorithms. Numerically, we demonstrate that our screening rule improves the efficiency of elastic net Cox proportional hazard model with a significant gain on a variety of simulated and real scenarios without any loss of accuracy.

**Semi-supervised Estimation of Event Rate with Doubly-censored Survival Data**

*Yang Wang1,* ♦*Qingning Zhou2, Tianxi Cai3 and Xuan Wang3.*

1Washington University in St. Louis， 2University of North Carolina at Charlotte， 3Harvard University

Electronic Health Record (EHR) has emerged as a valuable source of data for translational research. To leverage EHR data for risk prediction and subsequently clinical decision support, clinical endpoints are often time to onset of a clinical condition of interest. Precise information on clinical event times are often not directly available and requires labor-intensive manual chart review to ascertain. In addition, events may occur outside of the hospital system, resulting in both left and right censoring or often termed as double censoring. On the other hand, proxies such as time to the first diagnostic code are readily available yet with varying degrees of accuracy. Using error-prone event times derived from these proxies can lead to biased risk estimates while only relying on manually annotated event times, which are typically only available for a small subset of patients, can lead to high variability. This signifies the need for semi-supervised estimation methods that can efficiently combine information from both the small subset of labeled observations and a large size of surrogate proxies. While semi-supervised estimation methods have been recently developed for binary and right-censored data, no methods currently exist in the presence of doubly censoring. This paper fills the gap by developing a robust and efficient Semi-supervised Estimation of Event rate with Doubly-censored Survival data (SEEDS) by leveraging a small set of gold standard labels and a large set of surrogate features. Under mild regularity conditions, we demonstrate that the proposed SEEDS estimator is consistent and asymptotically normal. Extensive simulation results illustrate that SEEDS performs well in finite samples and can be substantially more efficient compared to the supervised counterpart. We apply the SEEDS procedure to estimate the age-specific survival rate of type 2 diabetes (T2D) using EHR data from Mass General Brigham (MGB).

**Empirical Likelihood Inference for the Panel Count Data with Informative Observation Process** *Faysal Satter1, Yichuan Zhao2 and* ♦*Ni Li3.* 1Data Analytics and Computational Intelligence, Lowe’s Companies, 2Georgia State University, 3Hainan Normal University.

Panel count data refer to interval-censored recurrent event data. Each study subject can only be observed at discrete time points, leading to knowledge about the total number of events occurring between observations. The observation times can be also different among subjects and carry important information about the underlying recurrent process. In our talk, an empirical likelihood (EL) method for panel count data with informative observation times is proposed. Based on the influence function, we formulate an empirical likelihood ratio for the vector of regression coefficients, and the Wilk's theorem is established. Simulation studies are carried out to compare the performance of empirical likelihood with normal approximation methods. Finally, the EL method is compared with existing approaches, utilizing an illustrative example drawn from a bladder cancer study.

**New Feature Screening Methods for Massive Interval-censored Failure Time Data**

♦*Huiqiong Li*, *Zhimiao Cao*, *Jianguo Sun*, *Niansheng Tang*

Screening important features has become one of the important tasks in statistical analysis and correspondingly, various screening procedures have been proposed for various types of studies or data including both complete and incomplete data. However, these methods would be computationally costly or even infeasible when one faces massive health databases with both high dimensionality and huge sample size, which have become increasingly popular for comparative effectiveness and safety studies of medical products. In this paper, we consider such a type of incomplete data, interval-censored failure time data, that have not be discussed before and propose two procedures with the use of distance correlation and orthogonal sampling as well as the jackknife debiased average technique. The proposed approaches can be easily implemented and their sure screening and rank consistency properties are established. Simulation studies demonstrate that the proposed methods work well for practical situations and they are applied to the SEER breast cancer data.

**Session 24CHI5: Advanced Statistical Methods for Bulk and Single-Cell Genomics**

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

♦*Ying Ma1, Xiang Zhou2*

*1Brown University，2University of Michigan*

Spatially resolved transcriptomics (SRT) studies are becoming increasingly common and large, offering unprecedented opportunities in mapping complex tissue structures and functions. Here, we present IRIS, a computational method designed to characterize tissue spatial organization in SRT studies through accurately and efficiently detecting spatial domains. IRIS uniquely leverages single-cell RNA-seq data for reference-informed detection of biologically interpretable spatial domains, integrating multiple SRT slices while explicitly considering correlations both within and across slices. We demonstrate the advantages of IRIS through in-depth analysis of six SRT datasets encompassing diverse technologies, tissues, species, and resolutions. In these applications, IRIS achieves significant accuracy gains (39%-1,083%) and speed improvements (4.6-666.0) in moderate-sized datasets, while representing the only method applicable for large datasets including stereo-seq and 10x Xenium. As a result, IRIS reveals intricate brain structures, uncovers tumor microenvironment heterogeneity, and detects structural changes in diabetes-affected testis, all at a speed and accuracy unachievable by existing approaches.

**Categorization of 31 computational methods to detect spatially variable genes from spatially resolved transcriptomics data**

♦*Jingyi Jessica Li*1

1University of California, Los Angeles

In the analysis of spatially resolved transcriptomics data, detecting spatially variable genes (SVGs) is crucial. Numerous computational methods exist, but varying SVG definitions and methodologies lead to incomparable results. We review 31 state-of-the-art methods, categorizing SVGs into three types: overall, cell-type-specific, and spatial-domain-marker SVGs. Our review explains the intuitions underlying these methods, summarizes their applications, and categorizes the hypothesis tests they use in the trade-off between generality and specificity for SVG detection. We discuss challenges in SVG detection and propose future directions for improvement. Our review offers insights for method developers and users, advocating for category-specific benchmarking.

**A systematic evaluation of highly variable gene selection methods for single-cell RNA-sequencing**

*Ruzhang Zhao*1, *Jiuyao Lu*1, *Weiqiang Zhou*1, *Ni Zhao*1, ♦*Hongkai Ji*1

1Johns Hopkins Bloomberg School of Public Health

Selecting highly variable genes (HVGs) or features (HVFs) is a key component of many scRNA-seq data analysis pipelines. Here we conduct a systematic benchmark study of 47 existing and new HVG selection methods using 19 benchmark datasets and an average of 18 evaluation criteria per method. We found that a hybrid approach robustly outperformed the other methods used individually. We developed an R package mixhvg that delivers these methods. Users can conveniently use this package to perform HVG selection independently or as part of their custom data analysis pipelines.

**Deconvoluting cell state distribution from bulk RNA-seq data**

*Liyang*1, *Xiwei Sun*2, *Ting Qi*2, ♦*Jian Yang*3

1Song Westlake University, 2Westlake University, 3Westlake University

Deconvoluting cell-state abundances from bulk RNA-seq data can add considerable value to existing data, but achieving fine-resolution and high-accuracy deconvolution remains a challenge. Here, we introduce MeDuSA, a mixed model-based method that leverages single-cell RNA-seq data as a reference to estimate cell-state abundances along a one-dimensional trajectory in bulk RNA-seq data. The advantage of MeDuSA lies primarily in estimating cell abundance in each state while fitting the remaining cells of the same type individually as random effects. Extensive simulations and real-data benchmark analyses demonstrate that MeDuSA greatly improves the estimation accuracy over existing methods for one-dimensional trajectories. Applying MeDuSA to cohort-level RNA-seq datasets reveals associations of cell-state abundances with disease or treatment conditions and cell-state-dependent genetic control of transcription. Our study provides a high-accuracy and fine-resolution method for cell-state deconvolution along a one-dimensional trajectory and demonstrates its utility in characterizing the dynamics of cell states in various biological processes.

**Session 24CHI29: Early Clinical Trial Development and Biomarker**

**Evaluation of Pharmacokinetics in Patients with Impaired Renal Function – Study Design, Data Analysis, and Impact on Dosing**

♦*Kong Xin*

According to ICH guideline E8 (R1), studies in special populations (e.g., cardiovascular disease, diabetes, hepatic and renal impairment) should be considered. For instance, renal impairment may alter the exposure-response relationship for a drug, especially in patients with severely impaired renal function. This presentation will provide a case on study design, statistical analysis and result interpretation in subjects with decreased renal function.

**Statistical considerations for non-oncology first-in-human clinical trials**

♦*Jiaqing Wang*

The purpose of the first-in-human (FIH) trials is to obtain safety, tolerability, and pharmacokinetic information on an investigational medication product (IMP). Traditionally, for non-oncology clinical trial, FIH clinical trial was most associate with a single ascending dose (SAD) study followed by multiple ascending dose (MAD) study. This presentation will briefly introduce the study design and related statistical analysis.

**Statistical considerations in the clinical drug interaction studies**

♦*Yanzhen Wu1*.

*1*Sanofi.

The severe drug-drug interaction (DDI) will cause morbidity or even modality after patients take more than one medications. During drug development, the clinical drug interaction studies are conducted to assess the pharmacokinetic interaction between the investigational medicinal product (IMP) and the medications selected from in preclinical in vitro and in vivo studieswhich could interaction with IMP. This presentation will mainly introduce the statistical considerations in clinical drug interaction studies, including study design, sample size calculation, analysis using linear mixed models, approaches to determine no-effect boundaries, and interpreting results.

**Biomarker analysis in clinical trial to assist the investigation of drug's mechanism of action**

♦*Meiyue Wang1.*

*1*Sanofi

Biomarker analysis can be in company with clinical trials in the early phase of drug development (phase 2) to help explaining the clinical results and further investigate the drug's mechanism of action in finer scope. Notably, biomarker analyses in participant subgroups (classified by genetic characteristics, concomitant medications, comorbidity disease, etc.) provide benefits in better understanding of the disease pathogenesis, drug signaling, development and/or validation of a bioassay method, and possibility to identify new drug targets or indications. A general pipeline is introduced for omics data analysis through a modified case study to present how the biomarker analysis started with unbiased common analysis and moved forward driven by clinical results.

**Session 24CHI48: New Developments in Modeling High-Dimensional and Complex Data**

**Conformal knock-off conditional independence test with double robustness**

♦*Baoying Yang, Jing Qin, Yukun Liu.*

**Normalized Power Prior Bayesian Inference and Computation**

*Zifei Han*1, *Min Wang*2, *Tianyu Bai*3, ♦*Keying Ye*2.

1University of International Business and Economics, China, 2University of Texas at San Antonio, USA, 3FDA, USA

The power prior is a widely adopted tool for moderating the influence of historical data in clinical trial design and analysis. This involves raising the likelihood function of historical data to a power parameter, representing the heterogeneity between historical and new studies. In a comprehensive Bayesian framework, an intuitive extension is to introduce a hyperprior for the power parameter, allowing the posterior distribution to capture the degree of similarity between historical and current data. To align with the likelihood principle, a normalized power prior has been proposed and a normalizing factor must be computed. Unfortunately, the calculation of this factor involves integrating a prior multiplied by a fractional likelihood, creating a computational burden, particularly for intricate models.

This presentation provides an exploration of the normalized power prior and presents an efficient importance sampling algorithm tailored for calculating the normalizing factor. Our approach bypasses the intricate calculations associated with repeated computations over various power parameter values during posterior sampling. Additionally, we introduce a streamlined posterior sampling algorithm, facilitating the incorporation of a randomly selected power parameter with adaptive borrowing capabilities in diverse models. The efficacy of our proposed method is demonstrated through extensive simulation studies and a real-world data example, highlighting notable numerical efficiency.

**A general spatial-temporal framework for short-term building temperature forecasting at arbitrary locations with crowdsourcing weather data**

*Reisa Widjaja*1, *Wenbo Wu*1, *Zhi Zhou*2, ♦*Renhao Sun*3, *Hannah Fontenot*4, *Bing Dong*4

1The University of Texas at San Antonio, 2Argonne National Laboratory, 3The University of Texas at San Antonio, 4Syracuse University

Weather forecasting has been a critical component to predict and control building energy consumption for better building energy management. Without accessibility to other data sources, the onsite observed temperatures or the airport temperatures are used in forecast models. In this paper, we present a novel approach by utilizing the crowdsourcing weather data from neighboring personal weather stations (PWS) to improve the weather forecast accuracy around buildings using a general spatial-temporal modeling framework. The final forecast is based on the ensemble of local forecasts for the target location using neighboring PWSs. Our approach is distinguished from existing literature in various aspects. First, we leverage the crowdsourcing weather data from PWS in addition to public data sources. In this way, the data is at much finer time resolution (e.g., at 5-minute frequency) and spatial resolution (e.g., arbitrary location vs grid). Second, our proposed model incorporates spatial-temporal correlation information of weather variables between the target building and a set of neighboring PWSs so that underlying correlations can be effectively captured to improve forecasting performance. We demonstrate the performance of the proposed framework by comparing to the benchmark models on temperature forecasting for a building located at an arbitrary location at San Antonio, Texas, USA. In general, the proposed model framework equipped with machine learning technique such as Random Forest can improve forecasting by 50% compares with persistent model and has 90% chance to outperform airport forecast in short-term forecasting. In a real-time setting, the proposed model framework can provide more accurate temperature forecasting results compared with using airport temperature forecast for most forecast horizon. Moreover, we analyze the sensitivity of model parameters to gain insights on how crowdsourcing data from the neighboring personal weather stations impacts forecasting performance. Finally, we implement our model in other cities such as Syracuse and Chicago to test the model’s performance in different landforms and climate types.

**Differentiable Neural Networks with Repu Activation: With Applications to Score Estimation and Isotonic Regression**

♦*Guohao Shen*1, *Yuling Jiao*2, *Yuanyuan Lin*3, *JianHuang*1

1The Hong Kong Polytechnic University, 2Wuhan University, 3The Chinese University of Hong Kong

We study the properties of differentiable neural networks activated by rectified power unit (RePU) functions. We show that the partial derivatives of RePU neural networks can be represented by RePUs mixed-activated networks and derive upper bounds for the complexity of the function class of derivatives of RePUs networks. We establish error bounds for simultaneously approximating C^s smooth functions and their derivatives using RePUactivated deep neural networks. Furthermore, we derive improved approximation error bounds when data has an approximate low-dimensional support, demonstrating the ability of RePU networks to mitigate the curse of dimensionality. To illustrate the usefulness of our results, we consider a deep score matching estimator (DSME) and propose a penalized deep isotonic regression (PDIR) using RePU networks. We establish non-asymptotic excess risk bounds for DSME and PDIR under the assumption that the target functions belong to a class of C^s smooth functions. We also show that PDIR achieves the minimax optimal convergence rate and has a robustness property in the sense it is consistent with vanishing penalty parameters even when the monotonicity assumption is not satisfied. Furthermore, if the data distribution is supported on an approximate low-dimensional manifold, we show that DSME and PDIR can mitigate the curse of dimensionality.

**Session 24CHI82: Recent Advances on High-Dimensional Inference and Generative Models**

**A leave-one-out approach to approximate message passing**

♦*Zhigang Bao1, Qiyang Han2, Xiaocong Xu1*

*1HKUST, 2Rutgers University*

In this talk, we will introduce a non-asymptotic, leave-one-out representation for the AMP iterate that holds under a broad class of random matrix models with general variance profiles. In contrast to the typical AMP theory that describes the empirical distributions of the AMP iterate via a low dimensional state evolution, our leave-one-out representation yields an intrinsically high dimensional state evolution formula which provides non-asymptotic characterizations for the possibly heterogeneous, entrywise behavior of the AMP iterate under the prescribed random matrix models.

Our leave-one-out method differs significantly from the widely adopted conditioning approach for rotational invariant ensembles, and relies instead on an inductive method that utilizes almost solely integration-by-parts and concentration techniques.

**SIMPLE-RC: Group Network Inference with Non-Sharp Nulls and Weak Signals**

*Jianqing Fan*1, *Yingying Fan*2, *Jinchi Lv*2, ♦*Fan Yang*3

1Princeton University, 2University of Southern California, 3Tsinghua University

The recent work of Fan, Fan, Han and Lv (2022) introduced a general framework of statistical inference on membership profiles in large networks (SIMPLE) for testing the sharp null hypothesis that a pair of given nodes share the same membership profiles. In real applications, there are often groups of nodes under investigation that may share similar membership profiles at the presence of relatively weaker signals than the setting considered in SIMPLE. To address these practical challenges, we propose a SIMPLE method with random coupling (SIMPLE-RC) for testing the non-sharp null hypothesis that a group of given nodes share similar (not necessarily identical) membership profiles under weaker signals. Utilizing the idea of random coupling, we construct our test as the maximum of the SIMPLE tests for subsampled node pairs from the group. Such technique reduces significantly the correlation among individual SIMPLE tests while largely maintaining the power, enabling delicate analysis on the asymptotic distributions of the SIMPLE-RC test. These new theoretical developments are empowered by a second-order expansion of spiked eigenvectors, built upon our work for random matrices with weak spikes. Based on joint work with Jianqing Fan, Yingying Fan and Jinchi Lv.

**SOFARI: High-Dimensional Manifold-Based Inference**

♦*Zemin Zheng*1, *Xin Zhou*1, *Yingying* *Fan*2, *Jinchi Lv*2

1USTC, 2USC

Multi-task learning is a widely used technique for harnessing information from various tasks. Recently, the sparse orthogonal factor regression (SOFAR) framework, based on the sparse singular value decomposition (SVD) within the coefficient matrix, was introduced for interpretable multi-task learning, enabling the discovery of meaningful latent feature-response association networks across different layers. However, conducting precise inference on the latent factor matrices has remained challenging due to orthogonality constraints inherited from the sparse SVD constraint. In this paper, we suggest a novel approach called high-dimensional manifold-based SOFAR inference (SOFARI), drawing on the Neyman near-orthogonality inference while incorporating the Stiefel manifold structure imposed by the SVD constraints. By leveraging the underlying Stiefel manifold structure, SOFARI provides bias-corrected estimators for both latent left factor vectors and singular values, for which we show to enjoy the asymptotic mean-zero normal distributions with estimable variances. We introduce two SOFARI variants to handle strongly and weakly orthogonal latent factors, where the latter covers a broader range of applications. We illustrate the effectiveness of SOFARI and justify our theoretical results through simulation examples and a real data application in economic forecasting.

**Approximate message passing: A non-asymptotic framework and beyond**

♦*Yuting Wei*1

1University of Pennsylvania

Approximate message passing (AMP) emerges as an effective iterative algorithm for solving high-dimensional statistical problems. However, prior AMP theory, which focused mostly on high-dimensional asymptotics, fell short of predicting the AMP dynamics when the number of iterations surpasses o(log n / log log n) (with n the problem dimension). To address this inadequacy, this talk introduces a non-asymptotic framework towards understanding AMP. Built upon a new decomposition of AMP updates in conjunction with well-controlled residual terms, we lay out an analysis recipe to characterize the finite-sample convergence of AMP up to O(n / polylog(n)) iterations. We will discuss concrete consequences of the proposed analysis recipe in the Z2 synchronization problem; more specifically, we predict the behavior of randomly initialized AMP for up to O(n/poly(\log n)) iterations, showing that the algorithm succeeds without the need of a careful spectral initialization and also a subsequent refinement stage (as conjectured recently by Celentano et al.)

**Session 24CHI115: Statistical Inference with Lean Assumption**

**BELIEF in Dependence: Leveraging Atomic Linearity in Data Bits for Rethinking Generalized Linear Models**

*Benjamin Brown1,* ♦*Kai Zhang1, Xiao-Li Meng2*

1UNC Chapel Hill, 2Harvard University

Two linearly uncorrelated binary variables must be also independent because non-linear dependence cannot manifest with only two possible states. This inherent linearity is the atom of dependency constituting any complex form of relationship. Inspired by this observation, we develop a framework called binary expansion linear effect (BELIEF) for understanding arbitrary relationships with a binary outcome. Models from the BELIEF framework are easily interpretable because they describe the association of binary variables in the language of linear models, yielding convenient theoretical insight and striking Gaussian parallels. With BELIEF, one may study generalized linear models (GLM) through transparent linear models, providing insight into how the choice of link affects modeling. For example, setting a GLM interaction coefficient to zero does not necessarily lead to the kind of no-interaction model assumption as understood under their linear model counterparts. Furthermore, for a binary response, maximum likelihood estimation for GLMs paradoxically fails under complete separation, when the data are most discriminative, whereas BELIEF estimation automatically reveals the perfect predictor in the data that is responsible for complete separation. We explore these phenomena and provide related theoretical results. We also provide preliminary empirical demonstration of some theoretical results.

**On partial envelope approach for modeling spatial-temporally dependent data**

♦*Zhigen Zhao*1, *Xin Xing*2

1Temple University, 2Virginia Tech University

In the new era of big data, modeling multivariate spatial-temporally dependent data is a challenging task due to the dimensionality of the features and complex spatial-temporal associations among the observations across different locations and time points. To improve the estimation efficiency, we propose a spatial-temporal partial envelope model which is parsimonious and effective in modeling high-dimensional spatial-temporal data. The partial envelope model was proposed under a linear coregionalization model framework which allows heterogenous spatial-temporal covariance structure for different components of the response vector. The maximum likelihood estimator for the proposed model can be obtained through a Grassmann manifold optimization. We obtained a complete asymptotic result for the estimator and conduct thorough empirical simulations to demonstrate the soundness and effectiveness of the proposed method. We also apply the proposed model to analyze the crowdsourcing weather data collected from personal weather stations in the city of San Antonio, TX of the United States.

**Controlling False Discovery Rate in High-Dimensional Linear Regression: The Gaussian Mirror Approach**

♦*Xin Xing*1, *Zhigen Zhao*2*，Jun Liu*3.

1Virginia Tech, 2Temple University, 3Harvard University.

Identifying key variables that influence outcomes in linear regression models, while also maintaining control over the false discovery rate (FDR), poses a significant challenge in statistical analysis. This presentation introduces the Gaussian Mirror (GM) method, a novel approach for identifying crucial variables in linear regression models while controlling the false discovery rate (FDR). By creating a pair of mirror variables for each predictor using Gaussian perturbations, the GM method improves variable selection. It's compatible with standard regression techniques like ordinary least-square and Lasso and offers the flexibility of applying mirror variables pre or post-selection. The key advancement of the GM method lies in its capacity to generate test statistics that effectively maintain the FDR at a predefined level under realistic covariate dependence assumptions. Our analysis showcases the GM method's superiority in managing FDR constraints, especially in situations with high covariate correlation and a dense array of influential variables. This presentation will cover the GM method's innovative approach, theoretical foundation, and empirical efficacy, offering attendees valuable insights into tackling complex statistical challenges.

**Session 24CHI114: Statistical Genetics: Using Statistical Tools to Elucidate the Genetic Basis of Complex Traits**

**Bayesian Inference of Local Genetic Correlation**

♦*Xiaoxuan Xia1, Ming Hin Ng2, Maggie Haitian Wang2, Lin Hou3, Yingying Wei2*

1Tianjin University, 2The Chinese University of Hong Kong, 3Tsinghua University

The genetic correlation quantifies the genetic similarity of two traits, thus having important applications in variant identification, risk prediction and causal inference. The overall genetic correlation measures the average correlation of genetic effects across the genome, whereas the local genetic correlation focuses on specific genomic regions. Although estimations of the overall genetic correlation have become routine for genome-wide association studies (GWAS), statistical methods that provide robust estimation of local genetic correlations are still lacking. On the one hand, genetic effects and hence the local genetic correlations are very heterogeneous across the genome. On the other hand, compared to the number of single nucleotide polymorphism (SNPs) which is on the scale of millions, the number of SNPs contributed to the local genetic correlation of a given region is very small. Here, we propose a statistical method, BAyesian inference of LOCal genetic correlation (BALOC), to provide robust estimation of local genetic correlations. Via a hierarchical model, BALOC is able to borrow information across the whole genome when performing region-specific inference. Simulation studies demonstrate that BALOC substantially outperforms the state-of-the-art method. Application of BALOC to GWAS data from the UK Biobank shows novel biological insights.

**Multi-omic Genetic Scores for Risk Prediction of Complex Human Diseases**

*Yafang Qiu*1, *Lin Hou*2, ♦*Nayang Shan*1.

1Capital University of Economics and Business, 2Tsinghua University

Polygenic risk score (PRS) has been a powerful approach in the risk prediction of complex human diseases. Many studies incorporated internal information, such as effect size distribution, or external information, such as linkage disequilibrium, functional annotation, and pleiotropy among multiple diseases, to optimize the performance of PRS. To leverage on multi-omic datasets, we proposed a framework to generate multi-omic genetic scores, in which plasma proteomics, serum metabolomics, messenger RNA expression levels were imputed and weighted for risk prediction. We evaluated the performance of multi-omic genetic scores in the risk prediction of complex traits of UK Biobank, which increase prediction accuracy over PRS across the tested traits.

**A Scalable and Accurate Analysis Framework to Control for Sample Relatedness in Large-scale Genome-wide Association Studies and Its Application to 79 Longitudinal Traits in UK Biobank**

♦*He Xu, Wenjian Bi*

Sample relatedness is a major confounder in genome-wide association studies (GWAS) and could result in the inflation of type one error rates if not appropriately controlled. A common strategy is to incorporate a random effect related to genetic relatedness matrix (GRM) into regression models. However, for a large-scale GWAS, it is challenging to apply this strategy to complex traits with complicated structures, such as longitudinal traits. In this work, we propose a scalable and accurate analysis framework, SPAGRM, which controlled the sample relatedness via the precise approximation of the joint distribution of genotypes. SPAGRM can utilize GRM-free models and thus is applicable to a wide variety of trait types and statistical methods. A hybrid strategy including saddlepoint approximation (SPA) greatly increases the accuracy to analyze low-frequency and rare genetic variants, especially when the phenotypic distribution is unbalanced. Due to the wide applicability, SPAGRM can conduct valid longitudinal trait GWAS based on proper analytical approaches and underlying assumptions. Meanwhile, we propose SPAGRM(CCT) to aggregate the results following different models via Cauchy combination test (CCT), which can serve as an optimal unified approach in a longitudinal trait GWAS. Extensive simulation studies and real data analyses demonstrated that SPAGRM-based approaches have well controlled type I error rates and can gain power for a longitudinal trait analysis. Expanding upon the previous studies, we implemented a refined and meticulous QC pipeline to extract 79 longitudinal traits from UK Biobank primary care data. SPAGRM identified 7,463 genetic loci for these traits, which is a pioneering attempt to conduct GWAS for a majority of these traits as a longitudinal phenotype.

**Leveraging Cross-population Fine-mapping to Strengthen cis-Mendelian Randomization**

♦*Mingxuan Cai*1

1City University of Hong Kong

By integrating GWASs and resources from expression quantitative trait loci (eQTL) mapping studies, cis-Mendelian randomization (cis-MR) seeks to determine the causal effect of gene expression on human complex traits. However, two key challenges have hampered the accurate identification of causal genes. First, eQTLs inherited at a low recombination rate often harbor multiple causal variants in linkage disequilibrium (LD), making it difficult to identify independent instrumental variables (IVs), and limiting the power of cis-MR. Second, eQTL variants can affect the outcome trait through pathways other than the target gene (i.e. horizontal pleiotropy), which violates the MR assumption and leads to false positive results. Here, we introduce a statistical method called XMR that leverages cross-population fine-mapping to identify causal SNPs of gene expression as IVs for MR analysis, which helps maximize the MR power. At the same time, we explicitly correct for the horizontal pleiotropy by using genome-wide information, effectively controlling the type-I error in cis-MR.

**Session 24CHI121: Statistical Methods in Causal Inference and Decision Making**

**Communication-Eﬀicient Precision Matrix Estimation by Distributed Refitted Cross-Validation**

*Qiao Zheng1, Riquan Zhang2,* ♦*Yan Zhong1.*

1East China Normal University, 2Shanghai University of International Business and Economics

This study develops a distributed refitted cross-validation (DRCV) method to estimate the precision matrix in distributed settings, whose communication complexity is linear to the number of non-zero entries in the precision matrix. The proposed method requires two rounds of communication. The first round selects non-zero positions of the precision matrix, while the second round gives an unbiased estimation for the selected positions. To address the overfitting issue regarding this inference-after-selection strategy, the proposed method utilizes the data-splitting technique. The authors establish the selection and estimation consistency of the proposed method and illustrate that the estimation error of the DRCV estimator achieves a minimax convergence rate under mild conditions. Numerical studies on simulated datasets and a real high-frequency stock dataset show the superior performance and efficiency of the proposed method. The proposed DRCV can be easily generalized to a variety of distributed parameter estimation problems where the parameter space exhibits a sparse structure.

**Bi-Level Offline Reinforcement Learning**

♦*Wenzhuo Zhou1.*

1University of California Irvine

We study offline reinforcement learning (RL) which seeks to learn a good policy based on a fixed, pre-collected dataset. A fundamental challenge behind this task is the distributional shift due to the dataset lacking sufficient exploration, especially under function approximation. To tackle this issue, we propose a bi-level structured policy optimization algorithm that models a hierarchical interaction between the policy (upper-level) and the value function (lower-level). The lower level focuses on constructing a confidence set of value estimates that maintain sufficiently small weighted average Bellman errors, while controlling uncertainty arising from distribution mismatch. Subsequently, at the upper level, the policy aims to maximize a conservative value estimate from the confidence set formed at the lower level. This novel formulation preserves the maximum flexibility of the implicitly induced exploratory data distribution, enabling the power of model extrapolation. In practice, it can be solved through a computationally efficient, penalized adversarial estimation procedure. Our theoretical regret guarantees do not rely on any data-coverage and completeness-type assumptions, only requiring realizability. These guarantees also demonstrate that the learned policy represents the best effort among all policies, as no other policies can outperform it.

**On Learning Necessary and Sufficient Causal Graphs**

♦*Hengrui Cai1.*

1University of California Irvine

The causal revolution has spurred interest in understanding complex relationships in various fields. Most existing methods aim to discover causal relationships among all variables in a large-scale complex graph. However, in practice, only a small number of variables in the graph are relevant for the outcomes of interest. As a result, causal estimation with the full causal graph -- especially given limited data -- could lead to many falsely discovered, spurious variables that may be highly correlated with but have no causal impact on the target outcome. In this paper, we propose to learn a class of necessary and sufficient causal graphs (NSCG) that only contains causally relevant variables for an outcome of interest, which we term causal features. The key idea is to utilize probabilities of causation to systematically evaluate the importance of features in the causal graph, allowing us to identify a subgraph that is relevant to the outcome of interest. To learn NSCG from data, we develop a score-based necessary and sufficient causal structural learning (NSCSL) algorithm, by establishing theoretical relationships between probabilities of causation and causal effects of features. Across empirical studies of simulated and real data, we show that the proposed NSCSL algorithm outperforms existing algorithms and can reveal important yeast genes for target heritable traits of interest.

**Time-varying Mediation Analysis for Incomplete Data with Application to DNA Methylation Study for PTSD**

♦*Kecheng Wei*1，*Fei Xue*2， *Qi Xu*3，*Yubai Yuan*4，*Yuexia Zhang*5，*Guoyou Qin*1，*Agaz H. Wani*6，*Derek E. Wildman*6， *Monica Uddin*6，*Annie Qu*3*.*

1Fudan University，2Purdue University, 3University of California Irvine，4Penn State University，5The University of Texas at San Antonio，6University of South Florida

DNA methylation (DNAm) has been shown to mediate causal effects from traumatic experiences to post-traumatic stress disorder (PTSD). However, the scientific question about whether the mediation effect changes over time remains unclear. In this paper, we develop time-varying structural equation models to identify cytosine-phosphate-guanine (CpG) sites where DNAm mediates the effect of trauma exposure on PTSD, and to capture dynamic changes in mediation effects. The proposed methodology is motivated by the Detroit Neighborhood Health Study (DNHS) with high-dimensional and longitudinal DNAm measurements. To handle the non-monotone missing DNAm in the dataset, we propose a novel Longitudinal Multiple Imputation (LMI) method utilizing dependency among repeated measurements, and employ the generalized method of moments to integrate the multiple imputations. Simulations confirm that the proposed method outperforms existing approaches in various longitudinal settings. In DNHS data analysis, our method identifies several CpG sites where DNAm exhibits dynamic mediation effects. Some of the corresponding genes have been shown to be associated with PTSD in the existing literature, and our findings on their time-varying effects could deepen the understanding of the mediation role of DNAm on the causal path from trauma exposure to PTSD risk.

**Session 24CHI123: Statistical Modeling for Complex Data**

**Unifying Nonparametric Inference for Time Series Data with Long-range Dependence**

♦*Jiancheng Jiang1.*

*1University of NC Charlotte*

This paper studies robust estimation of a multivariate regression function and its derivatives under longrange dependent errors. The resulting estimators are local M-estimators based on the local linear smoother. Under mild conditions, consistency and joint asymptotic distributions of the estimators are obtained. We show that the nature of the asymptotic distributions depends on the amount of smoothingness, that is, the estimators are asymptotically non-Gaussian distributed for large bandwidth and Gaussian distributed for small bandwidth, which extends the previous non-robust results (Masry and Mielniczuk, 1999) to the present situation. It is demonstrated that the asymptotic behaviors of the local Mestimators under long-range dependence are quite different from those under weak dependence (Jiang and Mack, 2001). Hence, there are breakdown formulas for the limiting distributions of the local Mestimators, which depend on the bandwidth conditions. This brings challenging to statistical inference for nonparametric regression models with long-range dependence. To address this difficulty, we propose a reweighting local M-estimation approach to constructing interval estimators of the regression function and its derivatives. It is shown that the proposed reweighting local M-estimator has a unique asymptotically normal distribution under different bandwidth conditions. This allows us to construct confidence intervals of the regression function and its derivatives via simulations. Our numerical results demonstrate nice performance of the proposed methodology in finite sample situations.

**Dynamic and Static Enhanced BIRCH for Functional Data Clustering**

♦*Youxi Luo*1, *Wang Li*1, *Maozai Tian*2*.*

1Hubei University of Technology, 2Renmin University of China

Accurate and efficient clustering of functional data is of utmost importance in the era of big data. However, the current research falls short in fully considering the high-order differentiability inherent in functional data. The effective integration of derivative information for clustering poses a significant challenge. This paper presents the Dynamic and Static Enhanced-BIRCH (DSE-BIRCH) algorithm, which characterizes the proximity between functional samples not only based on static distances but also by considering dynamic discriminability based on derivative functions. We introduce a novel matrix factorization-based approach that transforms constant features, extracted through functional principal component analysis, into derivative features. These two sets of features are subsequently fused to form global clustering features, which are utilized by an improved BIRCH algorithm for hierarchical clustering. Notably, the dynamic and static similarities between samples are concurrently weighted during the clustering process. Empirical results on publicly available datasets and simulated datasets demonstrate that DSE-BIRCH effectively captures dynamic information and exhibits superior

clustering performance. Further experiments involving noise and complexity attest to the algorithm's robustness and low-complexity characteristics, highlighting its potential for applications on large-scalf unctional data.

**State-varying Synthetic Control Method**

*Yuying Sun*1, ♦*Shaoxin Hong*2, *Yongmiao Hong*1, *Wendun Wang*3, *Xinyu Zhang1.*

1School of Economics and Management, University of Chinese Academy of Sciences, 2Center for Economic Research, Shandong University, 3Econometric Institute, Erasmus University Rotterdam

We propose a novel synthetic control method with a dynamic weighting scheme to evaluate the impacts of social policy. The basic idea is to utilize the interdependence between different control units in a panel dataset to create the counterfactuals locally. Unlike the existing literature, we allow the weights to change over state variables and thus it is expected to capture potentially nonlinear features in economics and finance. It is shown that the treatment-effect estimator is asymptotically optimal in the sense of achieving the lowest possible local squared prediction error. The rate of the selected weights converging to the optimal weights to minimizing the expected local quadratic loss is established. Simulations and empirical applications are conducted to evaluate the finite sample performance of the proposed method.

**A General Framework to Extend Sufficient Dimension Reductions to the Cases of the Mixture MultiVariate Elliptical Distributions**

*Wenjuan Li*1, *Hongming Pei*1, *Ali Jiang*1, ♦*Fei Chen1.*

1Yunnan University of Finance and Economics

In the sufficient dimension reduction (SDR), many methods depend on some assumptions on the distribution of predictor vector, such as the linear design condition (L.D.C.), the assumption of constant conditional variance, and so on. The mixture distributions emerge frequently in practice, but they may not satisfy the above assumptions. In this article, a general framework is proposed to extend various SDR methods to the cases where the predictor vector follows the mixture elliptical distributions, together with the asymptotic property for the consistency of the kernel matrix estimators. For illustration, the extensions of several classical SDR approaches under the proposed framework are detailed. Moreover, a method to estimate the structural dimension is given, together with a procedure to check an assumption called homogeneity. The proposed methodology is illustrated by simulated and real examples.

**Session 24CHI107: Statistical Advances in Modern Biomedical Applications: Robustness, Privacy, and Heterogeneity**

**Robust Divergence-based High Dimensional Heterogeneous Graphical Modeling**

♦*Rongzhuo Nie1.*

1University of Chinese Academy of Sciences

Existing graph-based methods for high-dimensional heterogeneous analysis are not suitable for data deviating from a Gaussian distribution. We introduce robust gamma divergence into high-dimensional mixture graph models to address the heterogeneous analysis problem of non-normally distributed data. Combining regularization techniques, we propose a robust high-dimensional mixture graph model based on gamma divergence. The Gaussian density processed through gamma divergence can effectively handle data deviating from a normal distribution, exhibiting robustness. We investigate the statistical properties of this model in high-dimensional scenarios, including the robustness of estimation, bounds on parameter estimation errors for the graph model, and variable selection consistency for mean and precision matrix parameters. Monte Carlo simulations are employed to comprehensively assess the model, considering aspects such as estimation effectiveness and robustness in handling anomalous data. Subsequently, empirical analysis are conducted on high-dimensional heterogeneous omics data.

**Two Head Neural Network based on Maximum Rank Correlation Loss**

♦*Huichao Li1.*

1University of Chinese Academy of Sciences

Recent years, deep learning has been received much attention due to powerful performance in many areas. Exploiting a large amount of labeled data to train is one typical reason of its success. In data sparsity problem, multi-task learning(MTL) is a good recipe by exploring useful information from other related learning tasks. We study two-head neural network including typical classification an regression tasks with high dimension input. Maximum Rank Correlation (MRC) is a measure of correlation between the ranking of true labels and fitted labels. It is invariant with respect to the functional forms of prediction and error distribution, which can deal with different scale of classification and regression tasks. We apply MRC loss, which is robust to error distribution, to two-head neural network. High computational cost and difficulty to adapt to regression task is discussed in this paper. Variable selection effect with high dimension input and model performance will be developed in theory.

**Summary Statistics-based Association Test for Identifying the Pleiotropic Effects with Set of Genetic Variants**

♦*Deiliang Bu*, *Xiao Wang* and *Qizhai Li*.

Traditional genome-wide association study focuses on testing one-to-one relationship between genetic variants and complex human diseases or traits. While its success in the past decade, this one-to-one paradigm lacks efficiency because it does not utilize the information of intrinsic genetic structure and pleiotropic effects. Due to privacy reasons, only summary statistics of current genome-wide association study data are publicly available. Existing summary statistics-based association tests do not consider covariates for regression model, while adjusting for covariates including population stratification factors is a routine issue. In this work, we first derive the correlation coefficients between summary Wald statistics obtained from linear regression model with covariates. Then, a new test is proposed by integrating three-level information including the intrinsic genetic structure, pleiotropy, and the potential information combinations. Extensive simulations demonstrate that the proposed test outperforms three other existing methods under most of the considered scenarios. Real data analysis of polyunsaturated fatty acids further shows that the proposed test can identify more genes than the compared existing methods.

**Identification of Spatially Variable Pathways for scRNA-seq Studies**

♦*Han Yan*1 and *Shihua Zhang*1.

1Institute of Mathematics and Systems Science, Chinese Academy of Sciences

Recent technological advances make it possible to measure gene expression with information on the spatial locations of cells. However, tools to analyze these spatial data are still needed. As a key step in the analysis of spatial data, identifying spatially variable pathways (SVPs) receives much more attention than before. In terms of this issue, we propose a novel SVP detection method and apply it to scRNA-seq datasets, which identifies SVPs, and reveal some new biological insights.

**Session 24CHI62: Recent Advances in Complex Network**

**A Latent Space Model for Weighted Keyword Co-occurrence Networks with Applications in Knowledge Discovery in Statistics**

*Yan Zhang*1*,* ♦*Rui Pan*2*, Xuening Zhu*3*, Kuangnan Fang*1 *and Hansheng Wang*4*.*

1Xiamen University, 2Xiamen University Central University of Finance and Economics, 3Fudan University, 4Peking University

Keywords are widely recognized as pivotal in conveying the central idea of academic articles. In this article, we construct a weighted and dynamic keyword co-occurrence network and propose a latent space model for analyzing it. Our model has two special characteristics. First, it is applicable to weighted networks; however, most previous models were primarily designed for unweighted networks. Simply replacing the frequency of keyword co-occurrence with binary values would result in a significant loss of information. Second, our model can handle the situation where network nodes evolve over time, and assess the effect of new nodes on network connectivity. We utilize the projected gradient descent algorithm to estimate the latent positions, and establish theoretical properties of the estimators. In the real data application, we study the keyword co-occurrence network within the field of statistics. We identify popular keywords over the whole period as well as within each time period. For keyword pairs, our model provide a new way to assess the association between them. Finally, we observe that the interest of statisticians in the emerging research areas is gradually growing in recent years.

**Distribution-Free Matrix Prediction Under Arbitrary Missing Pattern**

*Meijia Shao*1 *and* ♦*Yuan Zhang*1.

1Ohio State University

This paper studies the open problem of conformalized entry prediction in a row/column-exchangeable matrix. The matrix setting presents novel and unique challenges, but there exists little work on this interesting topic. We meticulously define the problem, differentiate it from closely related problems, and rigorously delineate the boundary between achievable and impossible goals. We then propose two practical algorithms. The first method provides a fast emulation of the full conformal prediction, while the second method leverages the technique of algorithmic stability for acceleration. Both methods are computationally efficient and can effectively safeguard coverage validity in presence of arbitrary missing pattern. Further, we quantify the impact of missingness on prediction accuracy and establish fundamental limit results. Empirical evidence from synthetic and real-world data sets corroborates the superior performance of our proposed methods.

**Open-Set Graph Domain Adaptation via Separate Domain Alignment**

♦*Pengsheng Ji*1.

1University of Georgia

Domain adaptation has become an attractive learning paradigm, as it can leverage source domains with rich labels to deal with classification tasks in an unlabeled target domain. A few recent studies have developed domain adaptation approaches for graph-structured data. In the case of node classification task, current domain adaptation methods only focus on the closed-set setting, where source and target domains share the same label space. A more practical assumption is that the target domain may contain new classes that are not included in the source domain. Therefore, in this paper, we introduce a novel and challenging problem for graphs, i.e., open-set domain adaptive node classification, and propose a new approach to solve it. Specifically, we develop an algorithm for efficient knowledge transfer from a labeled source graph to an unlabeled target graph under a separate domain alignment (SDA) strategy, in order to learn discriminative feature representations for the target graph. Our goal is to not only correctly classify target nodes into the known classes, but also classify unseen types of nodes into an unknown class. Experimental results on real-world datasets show that our method outperforms existing methods on graph domain adaptation.

**The impact of uncertainty to humanitarian relief operations resilience and responsiveness**

*Yingying Sun*1, ♦*Ping Wang*2*.*

1Renmin University, 2James Madison University

This study investigates the impact of uncertainty and preparation on humanitarian relief operations, where resilience mediates the relationship between uncertainty/preparation and response. The sample of 1306 valid responses collected in China showed the importance of resilience of humanitarian relief operations as the total indirect effect size was much stronger than the total direct effect size. A preliminary analysis of a moderated mediation model with uncertainty and respondents job characteristics as moderators presents interesting results. While the direct effect size of preparation on responsiveness and resilience was between small and medium, the effect size of resilience on responsiveness was very large, which indicates the importance of increased resilience. Small effects are observed for job characteristics on responsiveness, and the interaction effect of job characteristics and preparation on resilience. A very large simple effect is observed for uncertainty on resilience, even if the interactive effect seems much smaller. The theoretical and practical implications of the findings will be presented along the future research directions.

**Session 24CHI79: Recent Advances in Statistical Modeling and Computation in Applications with Complex Data Structures**

**Discounting Effect Size When Borrowing External Data in Clinical Studies**

♦*Zhuanzhuan Ma1, Chul Ahn2, Bin Wang3 and Xuefeng Li3.*

1University of Texas Rio Grande Valley， 2Edwards Lifesciences， 3Food and Drug Administration

When borrowing information from external data to augment a current trial, many available methods discount the sample size but retain the effect size from previous studies. Discounting the sample size is just one way to discount the prior information. It may not be appropriate if the underlying assumption of unbiased treatment effect does not hold, for example, when the treatment effect in the historical study is likely higher than the one expected in the current trial. To tackle this potential issue, we study some methods to shrink the effect size from previous studies assuming that the prior effect size is higher than the true effect size. These methods include weighted mean, Bayesian hierarchical method for an individual study/subgroup, multiplicity adjusted mean, and dynamic/conditional shrinkage when borrowing information from external data for a patient population of interest. We evaluate the performance of these methods for normal and binomial endpoints through Monte Carlo simulation studies and compare them with some available methods of borrowing external data. Numerical results demonstrate that the proposed multiplicity adjustment method has good performance in terms of bias, type I error and power control. The Bayesian hierarchical modelling method has comparable performance as the multiplicity method when the hyper variance is well chosen or estimated. Some potential application scenarios are discussed and illustrated via a few hypothetical case studies. This research will add to the toolbox of utilizing external data to generate clinical evidence in clinical studies.

**Metamodeling in Computer Experiment for Quality Design: The Perspective of Statistical Test**

♦*Linhan Ouyang1 and Baoping Tao1.*

1Nanjing University of Aeronautics and Astronautics

Based on the perspective of statistical test, a computer metamodeling technique in the framework of Bayesian Kriging is proposed for quality design to tackle the uncertainty of quality variables and model structure. The proposed method not only screens out the significant variables for the Kriging mean function by applying the Bayesian hierarchical method, but also ensures the validity of the metamodel from the statistical perspective with multiple hypothesis test. In addition, parameter Bootstrap is applied to correct the underestimated variance in the construction of statistic. Firstly, a binary variable indicator is introduced for each candidate variable in the Kriging model to account for the factorial effect principles. After that, the Markov Chain Monte Carlo simulation is applied to identify the significant variables and estimate the posterior probability of each candidate model. Thirdly, multiple hypothesis testing and Bootstrap prediction variance correction are conducted to analyze the validity of the candidate models. Then, the optimal model is determined by evaluating the validity and generalization ability of the model. Finally, the result of the simulation demonstrates that the proposed approach maintains high predictive accuracy and robustness under different significance levels and sample sizes, and it is also verified through two industrial cases that our method can effectively select the significant factors under differential posterior probability and non-differential posterior probability scenarios.

**A Simple Two-Sample Bayesian Hotelling’s T2 for Equality of Means**

♦*Min Wang1.*

1The University of Texas at San Antonio

In this presentation, we introduce an explicit closed-form Bayes factor for comparing the means of two-sample populations. The proposed approach not only encompasses many existing Bayesian tests as special cases but also serves as a Bayesian counterpart to the classical Hotelling’s T2 test, highlighting its appealing properties in practical applications. It relies solely on data through Hotelling’s T2 test and can be easily integrated into a multivariate analysis course with a focus on Bayesian thinking. We also delve into the elicitation of prior specifications, a crucial step in the process. In addressing common challenges in high-dimensional mean vector testing problems, we discuss the generalization of this procedure in high-dimensional settings. Finally, we illustrate the practicality of this approach through several examples.

**A New Residual Subsampling Method for Skew-normal Mode Regression Model with Massive Data**

♦*Liucang Wu*

With the advent of big data, the fields of biomedicine and economics generate massive data with skew characteristics. Numerous methods have been proposed for modeling either skewed or massive data, whereas most existing methods cannot allow a direct handling of massive and skewed data. We first investigate the subsampling algorithms for skew-normal mode regression model, which include uniform subsampling, leverage subsampling, optimal subsampling, and vector mode subsampling. Since the aforementioned algorithms mainly leverage the value of the information module to calculate the sampling probability without accounting for the residuals in the modeling process. This observation motivates us to propose a novel residual subsampling method with applications to massive data. We then employ the signal-to-noise ratio (SNR) to carry out simulation studies to compare the performance of various sampling methods under various information quantities. Finally, a real-data example is provided for illustrative methods.

**Session 24CHI68: Recent Advances in Modeling Spatial Transcriptomics Data**

**SDEvelo: A Deep Generative Approach for TranScriptional Dynamics with Cell-specific Latent Time and Multivariate Stochastic Modeling**

*Xu Liao*1*,* ♦*Jin Liu*1

1The Chinese University of Hong Kong

Recently, RNA velocity has driven a paradigmatic change in single-cell RNA sequencing (scRNA-seq) studies, allowing the reconstruction and prediction of directed trajectories in cell differentiation and state transitions. However, most existing methods use dynamic modeling via ordinary differential equations (ODE) for individual genes in sequence and can lead to erroneous results, as they are inadequately able to fully capture the intrinsically stochastic nature of transcriptional dynamics governed by a cell-specific latent time across multiple genes. Here, we present SDEvelo, a novel deep generative approach to inferring RNA velocity by modeling the dynamics of unspliced and spliced RNAs via multivariate stochastic differential equations (SDE). Uniquely, SDEvelo explicitly models inherent uncertainty in transcriptional dynamics while estimating a cell-specific latent time across genes. Using both simulated and four scRNA-seq and spatial transcriptomics datasets, we show that SDEvelo can model the random dynamic patterns of mature-state cells while accurately detecting carcinogenesis. Additionally, the estimated gene-shared latent time can facilitate many downstream analyses for biological discovery. We demonstrate that SDEvelo is computationally scalable and applicable to both scRNA-seq and sequencing-based spatial transcriptomics data.

**Statistical Identification of Cell Type-specific Spatially Variable Genes in Spatial Transcriptomics**

*Lulu Shang, Peijun Wu and* ♦*Xiang Zhou*

An essential task in spatial transcriptomics involves identifying genes with spatial expression patterns, known as spatially variable genes (SVGs). Importantly, a subset of SVGs displays diverse spatial expression patterns within a given cell type, thus representing key transcriptomic signatures underlying cellular heterogeneity. Here, we present Celina, a statistical method for systematically detecting this subset of cell type-specific SVGs (ct-SVGs). Celina utilizes a spatially varying coefficient model to accurately capture each gene's spatial expression pattern in relation to the distribution of cell types across tissue locations, ensuring effective type I error control and high statistical power. We evaluated the performance of Celina through comprehensive simulations and applications to five real datasets, where we also adapted and examined existing methods originated from other analytic settings to detect ct-SVGs. Celina proves powerful compared to these ad hoc method adaptations in single cell resolution spatial transcriptomics and stands as the only effective solution for spot resolution spatial transcriptomics. In the real data applications, Celina uncovers ct-SVGs associated with tumor progression and patient survival in lung cancer, identifies metagenes with unique spatial patterns linked to cell proliferation and immune response in kidney cancer, and detects genes preferentially expressed near amyloid-β plaques in an Alzheimer's model. The ct-SVGs detected by Celina also enable novel biologically informed downstream analyses, unveiling functional cellular heterogeneity at an unprecedented scale.

**Resolving Tissue Complexity By Multi-modal Spatial Omics Modeling with MISO**

♦*Mingyao L*i1 *and Kyle Coleman*1.

1University of Pennsylvania.

Modeling multi-modal spatial omics data is crucial for understanding tissue complexities. MISO is a robust and versatile algorithm for feature extraction and spatial clustering, capable of integrating multiple modalities from diverse spatial omics experiments. Its effectiveness is demonstrated across various datasets, encompassing gene expression, protein expression, epigenetics, metabolomics, and imaging modalities. MISO outperforms two-modality methods in identifying biologically relevant spatial domains, representing a significant advancement in multi-modal spatial omics analysis.

**June 30 10:00-11:40**

**Session 24CHI13: Advances in Statistical Methods For-Omics Data Analysis**

**Cluster Analysis of Longitudinal Profiles for Compositional Count Data**

*Chenyang Duan*1*,* ♦*Yuan Jiang*2

1AbbVie， 2Oregon State University

To classify biological roles of different species in an ecological system, modern studies collect longitudinal and compositional counts of DNA sequences of taxonomically diagnostic genetic markers to measure the abundance of species over time. The major challenges of conducting this analysis are twofold: how to accommodate the complex dependence in this data type and how to model the longitudinal trajectories of the species' abundances. In this paper, we propose a novel method named COMPARING to cluster longitudinal profiles for compositional count data to address these challenges. In COMPARING, generalized estimating equation is used to account for both the compositional and longitudinal dependence structures, nonparametric B-spline approximation is used to model the longitudinal curves, and a pairwise-distance penalization is used to identify subgroups with similar longitudinal patterns. We establish the convergence rate of the estimated curves and conclude that the true subgroups can be correctly identified with a high probability. We also conduct simulation studies to show the advantage of COMPARING over its competitors in clustering longitudinal trajectories from compositional count data. Finally, we apply COMPARING to study the co-existence of blood-borne parasites in African buffalo and demonstrate how the method successfully detects biologically meaningful subgroups of parasites for competition-colonization trade-off.

**mbDecoda: A Debiased Approach to Compositional Data Analysis for Microbiome Surveys**

♦*Tao Wang*1

1Shanghai Jiao Tong University

Potentially pathogenic or probiotic microbes can be identified by comparing their abundance levels between healthy and diseased populations, or more broadly, by linking microbiome composition with clinical phenotypes or environmental factors. However, in microbiome studies, feature tables provide relative rather than absolute abundance of each feature in each sample, as the microbial loads of the samples and the ratios of sequencing depth to microbial load are both unknown and subject to considerable variation. Moreover, microbiome abundance data are count-valued, often over-dispersed, and contain a substantial proportion of zeros. To carry out differential abundance analysis while addressing these challenges, we introduce mbDecoda, a model-based approach for debiased analysis of sparse compositions of microbiomes. mbDecoda employs a zero-inflated negative binomial model, linking mean abundance to the variable of interest through a log link function, and it accommodates the adjustment for confounding factors. To efficiently obtain maximum likelihood estimates of model parameters, an Expectation Maximization algorithm is developed. A minimum coverage interval approach is then proposed to rectify compositional bias, enabling accurate and reliable absolute abundance analysis. Through extensive simulation studies and analysis of real-world microbiome datasets, we demonstrate that mbDecoda compares favorably to state-of-the-art methods in terms of effectiveness, robustness, and reproducibility.

**BFAST：Joint Dimension Reduction and Spatial Clustering with Bayesian Factor Analysis for Zero-inflated Spatial Transcriptomics Data**

♦*Yang Xu*1, *Dian Lv*1, *Xin Zhao*1

1BGI-research

The development of spatially resolved transcriptomics (ST) technologies has made it possible to the measure gene expression profiles coupled with cellular spatial context and assist biologists comprehensively characterize cellular phenotype heterogeneity and tissue microenvironment. Spatial clustering is vital for biological downstream analysis. However, due to high noise and dropout events, clustering spatial transcriptomics data poses numerous challenges due to the lack of effective algorithms. Here we develop a novel method, jointly performing dimension reduction and spatial clustering with Bayesian Factor Analysis for zero-inflated Spatial Transcriptomics data (BFAST). BFAST has showcased exceptional performance on simulation data and real spatial transcriptomics datasets, as proven by benchmarking against currently available methods. It effectively extracts more biologically informative low-dimensional features compared to conventional dimension reduction methods, thereby enhancing clustering performance. Moreover, it holds the promise to improve trajectory inference and visualization for downstream analyses of cellular trajectories.

**Large-scale Optimal Transport and Its Application in Biomedical Research**

♦*Jingyi Zhang*1.

1Tsinghua University.

Optimal transport has been one of the most exciting subjects in mathematics, starting from the 18th century. As a powerful tool to transport between two probability measures, optimal transport methods have been nowadays in a remarkable proliferation of modern data science applications. To meet the big data challenges, various computational tools have been developed in the recent decade to accelerate the computation for optimal transport methods. In this talk, we present a projection-based optimal transport method. We then discuss its real-world applications in biomedical research.

**Session 24CHI36: High Dimensional Statistical Inference and Its Applications**

**ATM: An Aggregation Test of Moments Approach for Assessing High-Dimensional Normality**

*Hengjian Cui*1*,* ♦*Lingyue Zhang*2

1Capital Normal University， 2Dongbei University of Finance and Economics

The Gaussian assumption is the most common and widely used distribution in statistical methodology. The various affine invariant tests for normality rely on the inverse of the covariance matrix and do not apply to high-dimensional data. The Gaussian moments are equal to the specific values, which implies an exclusive quantitative relationship. This paper introduces two novel indices, the co-third moment and the co-fourth moment, to characterize the shape of the distribution relative to the high-dimensional normal distribution. Using the relationship among Gaussian moments, two new test statistics for high-dimensional normality are proposed, which substantially avoid using the inverse of the covariance matrix and are easy to implement. By aggregating the strengths of the two tests and using the power enhancement technique, this study develops a more powerful test of high-dimensional normality. Finally, we conduct numerical studies and real data analysis to verify the superior power performance of our method with a well-controlled size, compared to other existing tests, and illustrate the capability of distinguishing between Gaussian distribution and alternatives in both low-dimensional and high-dimensional cases.

**Confidence Intervals for High-dimensional Regression Using Model Selection and Model Averaging**

♦*Wenwen Guo*1, *Yixuan Yin*1

1Capital Normal University

This paper presents a new method for construction of a model-averaged confidence intervals in ultra-high dimensional linear models, based on the idea of model averaging tail areas of the sampling distributions of the single-model estimates. We propose a two-stage procedure based on a random data-splitting strategy to establish the confidence intervals. In the first stage, we devide the data randomly into two parts and apply the refitted and cross validation procedure to identify a candidate model. In the second stage, we use a model-averaged method to construct the cofidence intervals, where AIC weights or BIC weights are considered. We demonstrate the excellent finite-sample performance of the proposed method using Monte Carlo simulations and a real-data example, in terms of coverage rate and interval width.

**Cross Projection Test for High-dimensional Mean Vectors**

♦*Guanpeng Wang*1, *Hengjian Cui*2

A cross projection test (CPT) technique for a one-sample vector in a high-dimensional setting is introduced. To overcome the problems caused by the curse of dimensionality, we construct test statistics by employing a projection test to project high-dimensional samples into one-or multi-dimensional directions. First, we randomly split the sample into two groups. We then find the p projection directions from a sample covariance matrix of the first group of samples. The second group is used to construct a projection statistic and perform the test. Second, we find the projection directions by exchanging the order of the two groups of samples, and we perform the test again to obtain another test statistic. Finally, we construct the CPT statistic by adding the two asymptotically uncorrelated test statistics together using the cross projection technique, such that the information from the two independent split samples can be fully utilized. The simulation results show that our proposed cross projection test controls the type I error well, and it is more powerful than the existing mean tests for some covariance matrix structures. Meanwhile, after applying the power enhancement technique, the CPT method performs non-trivially in general cases, especially for testing against sparse alternatives. A real gene-data analysis illustrates that the performance of our CPT is quite well.

**A Distance and Kernel-based Framework for Global and Local Two-sample Conditional Distribution Testing**

♦*Zhouxi Li*1, *Jian Yan*2 and *Xianyang* *Zhang*3.

1Xiamen University, 2Cornell University, 3Texas A&M University

Testing for the equality of two conditional distributions is critical in numerous modern applications such as transfer learning and program evaluation. However, this fundamental problem has surprisingly received little attention in the literature. The primary objective of this paper is to establish a distance and kernel-based framework for two-sample conditional distribution testing that is adaptable to multivariate distributions and allows for heterogeneity in the marginal distributions. We propose two metrics, the conditional generalized energy distance and the conditional maximum mean discrepancy, which completely characterize the homogeneity of two conditional distributions. Utilizing these metrics, we develop both local and global tests that can identify local and global discrepancies between two conditional distributions. In theory, we derive the convergence rates as well as the asymptotic distributions of the local and global tests under both the null and alternative hypotheses. To approximate the finite-sample distributions of the test statistics, we employ a novel local bootstrap procedure. Our proposed local and global two-sample conditional distribution tests demonstrate reliable performance through simulations and real data analysis.

**Session 24CHI63: Recent Advances in Data Integration and Structure Identification**

**Orthogonal Common-source and Distinctive-source Decomposition Between High-dimensional Data Views**

♦*Hai Shu1, Hongtu Zhu2*

*1New York University， 2University of North Carolina at Chapel Hill*

Modern biomedical studies often collect multi-view data, that is, multiple types of data measured on the same set of objects. A typical approach to the joint analysis of two high-dimensional data views/sets is to decompose each data matrix into three parts: a low-rank common-source matrix that captures the shared information across data views, a low-rank distinctive-source matrix that characterizes the individual information within each single data view, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common-source and distinctive-source matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive-source matrices. The latter guarantees that no more shared information is extractable from the distinctive-source matrices. We propose a novel decomposition method that defines the common-source and distinctive-source matrices from the L2 space of random variables rather than the conventionally used Euclidean space, with careful construction of the orthogonal relationship between distinctive-source matrices. The proposed estimators of common-source and distinctive-source matrices are shown to be asymptotically consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and real data analysis.

**Estimating Interpretable Heterogeneous Survival Treatment Effect Under Counterfactual Framework**

*Na Bo*1, ♦*Ying Ding*1

1University of Pittsburgh

Estimating heterogeneous treatment effects plays a central role in precision medicine as it provides informative guidelines in tailoring existing therapies for each patient to get the optimal treatment. Recently, meta-learning approaches have received a lot of attention in estimating conditional average treatment effect (CATE) by using multi-step algorithms coupled with flexible machine learning methods. In this work, we provide a meta-learning framework to estimate CATE on survival outcomes. We address the advantages of adapting existing meta-learning methods to survival outcomes through comprehensive simulations under different randomized clinical trials (RCT) and observational study settings and provide guidelines for implementing these methods to survival outcomes. Beyond black box predictions, we seek interpretable CATE predictions by evaluating meta-learning algorithms in terms of identifying prognostic and predictive biomarkers and identifying interpretable subgroups. We also apply the proposed methods to an RCT and an observational study to estimate CATE and make treatment recommendations.

**Integrative Nearest Neighbor Classifiers for Block-missing Multi-modal Data**

♦*Guan Yu*1

1University of Pittsburgh

Classifiers leveraging multi-modal data often have excellent classification performance. However, in certain studies, due to various reasons, some modalities are not collected from a sizable subset of participants and thus all data from those modalities are missing completely. Considering classification problems with a block-missing multi-modal training data set, we develop a new integrative nearest neighbor (INN) classifier. INN harnesses all available information in the training data set and the feature vector of the test data point effectively to predict the class label of the test data point without deleting or imputing any missing data. Given a test data point, INN determines the weights on the training samples adaptively by minimizing the worst-case upper bound on the estimation error of the regression function over a convex class of functions. As a weighted nearest neighbor classifier, INN suffers from the curse of dimensionality. Therefore, in high-dimensional scenarios, we propose a two-step INN, assuming that the regression function depends on features via sparse linear combinations of features. Our two-step INN estimates those linear combinations first, and then use them as new features to build the classifier. The effectiveness of our proposed methods have been demonstrated by both theoretical and numerical studies.

**High-Resolution Feature Identification on Data Heterogeneity in High-Dimensional Clustering**

♦*Lyuou Zhang*1，*Lulu Wang*2，*Xiwei Tang*3，*Wen Zhou*4

1School of Statistics and Management, Shanghai University of Finance of Economics， 2Gilead Sciences，3Department of Statistics, University of Virginia， 4Department of Biostatistics, School of Global Public Health,New York University

Simultaneously identifying heterogeneous subgroups and the informative features defining them, especially in the absence of responses and with a plethora of features, has long been a challenge in various domains, including omics studies, clinical research, and policy evaluation. Existing methods have either focused narrowly on global informative features or performed feature selection and group recovery as separate tasks, overlooking their interactions. Such methods might miss scientifically relevant information, and lead to suboptimal solutions to both feature identification and subgroup recovery. To overcome these limitations, we introduce a novel unsupervised learning approach, PAirwise REciprocal fuSE (PARSE), which concurrently pinpoints cluster-specific informative features and conducts high-dimensional clustering. Our method employs a new regularization that heavily penalizes features with minor differences across clusters, thus avoiding the selection of less informative features that define clusters. The oracle property of PARSE is obtained, and we establish lower bounds for both clustering and cluster-specific feature identification, affirming our method's optimality in both aspects. For implementations, we have devised an enhanced Expectation-Maximization algorithm, which is computationally feasible. Extensive numerical studies showcase PARSE's superiority over existing methods. In an application involving the identification of gene signatures in different subtypes of human pancreatic cells using single-cell RNAseq data, PARSE outperforms most mainstream methods in terms of identifying both the cell subtypes and corresponding gene signatures.

**Session 24CHI104: Spatial Data and Semiparametrics**

**Multivariate Spatial Autoregressive Models with Nonlinear Transformation: Specification, Identification and Estimation**

♦*Kai Yang*1, *Lung-fei Lee*1

1Shanghai University of Finance and Economics

This paper investigates multivariate spatial autoregressive (SAR) models with nonlinear transformation, which extend the univariate SAR model with nonlinear transformation functions and linear multivariate SAR models. We propose two stage least square, three stage least square, generalized method of moments and maximum likelihood estimation methods and study the asymptotic properties of the estimators. We also design Monte Carlo experiments to investigate the nite sample performance of the estimators under different model specifications.

**Matrix-valued Network Autoregression Model with Latent Group Structure**

♦*Yimeng Ren1, Xuening Zhu1, Ganggang Xu2, Yanyuan Ma3*

1Fudan University 2University of Miami 3The Pennsylvania State University

Matrix-valued time series data are frequently observed in a broad range of areas and have attracted great attention recently. In this work, we model network effects for high dimensional matrix-valued time series data in a matrix autoregression framework. To characterize the potential heterogeneity of the subjects and handle the high dimensionality simultaneously, we assume that each subject has a latent group label, which enables us to cluster the subject into the corresponding row and column groups. We propose a group matrix network autoregression (GMNAR) model, which assumes that the subjects in the same group share the same set of model parameters. To estimate the model, we develop an iterative algorithm. Theoretically, we show that the group-wise parameters and group memberships can be consistently estimated when the group numbers are correctly or possibly over-specified. An information criterion for group number estimation is also provided to consistently select the group numbers. Lastly, we implement the method on a Yelp dataset to illustrate the usefulness of the method.

**Nonparametric Network Vector Autoregression**

♦*Zixin Yang*1, *Xiaojun Song*1, *Jihai Yu*1

1Guanghua School of Management, Peking University

We consider network vector autoregression (NAR) for large-scale social networks with continuous response observed for each node at equally spaced time points. Our model allows the network and momentum effects to be nonlinear and nonparametric. For the model estimation, we propose a sieve least squares estimator and establish both its consistency and asymptotic normality. We also propose a nonparametric specification test for the linearity of the network and momentum effects. Under the null hypothesis of linearity, we show that the test statistic is asymptotically distributed as normal. Monte Carlo simulations show that the proposed estimators and tests perform well in finite samples.

**Shrinkage Estimation for Time Varying High-order Spatial Autoregressive Models**

♦Jin Liu1, Yingqiu Zhu2, Xuening Zhu3, Hansheng Wang4

1Nankai University, 2University of International Business and Economics， 3Fudan University， 4Peking University

The spatial autoregressive (SAR) model and its various extensions have been widely used to model data with spatial or network dependence. In this article, we propose a novel time varying SAR model with multiple weight matrices to model different types of spatial dependence and the complex relationships between the response and covariates. Accordingly, we present a local maximum likelihood estimator (LMLE). Moreover, an one-step updating estimator (one-step LMLE) is also proposed to reduce computational cost. The asymptotic properties of both LMLE and one-step LMLE are established. Furthermore, we develop two variable selection methods to simultaneously select significant covariates and weight matrices. We prove that the proposed two methods are selection consistent and enjoy oracle property under certain regularity conditions. Correspondingly, two novel Bayesian information criterions (BIC) are proposed for tuning parameter selection, respectively. Simulation studies and an empirical example are presented to illustrate the usefulness of the proposed time varying SAR model and variable selection procedures.

**Session 24CHI109: Statistical and Computational Advances for Complex Biological Data Analysis**

**Multi-marker Genetic Association and Interaction Tests Based on the Accelerated Failure Time Model**

♦*Chenxi Li1, Di Wu1, Qing Lu2*

1Michigan State University， 2University of Florida

Kernel-based multi-marker tests for survival outcomes use primarily the Cox model to adjust for covariate. The proportional hazards assumption made by the Cox model could be unrealistic, especially in the long-term follow-up. We develop a suite of novel multi-marker survival tests for genetic association and interaction based on the accelerated failure model, which is a popular alternative to the Cox model due to it's direct physical interpretation. The tests are based on the asymptotic distributions of their test statistics and are thus computationally efficient. The association tests can account for the heterogeneity of genetic effects across sub-populations/individuals to increase the power. All the new tests can deal with competing risks and left truncation. Moreover, we develop small-sample corrections to other tests to improve their accuracy under small samples. Extensive numerical experiments show that the new tests perform very well in various scenarios. An application to a genetic dataset of Alzheimer's disease illustrates the tests' practical utility.

**Integrative Modeling of Multi-omics Information in 3D Chromatin Reveals Complex Disease Mechanisms**

*Hao Wang*1, *Jiaxin Yang*1, *Yu Zhang*1, *Jianliang Qian*1, ♦*Jianrong Wang*1

1Michigan State University

One of the fundamental challenges in disease genetics is to delineate the molecular mechanisms linking the disruptive effects of genetic variants to complex phenotypes, especially for non-coding SNPs that are located distal to target genes. While the spatial chromosomal conformations in 3D space have been demonstrated to play pivotal roles in modulating long-range genetic associations, current experimental data of chromatin contacts (e.g. Hi-C and Capture-C) are highly limited to a few cellular-contexts. The data have extremely high rates of missing data, making high-resolution characterization of detailed genome folding challenging. Excessive missing data of chromatin contacts also makes the interpretation of long-range genetic associations infeasible. Here we developed a family of computational models, FLAMINGO and its variants, which are able to reconstruct 1kb-resolution spatial configurations of the human genome (i.e. the highest resolution to date) across diverse cell types and even at single-cell specific levels. FLAMINGO consistently demonstrates superior accuracy, orders-of-magnitude boost in scalability, and strong robustness against high rates of missing contacts. Integrative analysis of reconstructed 3D genome structures with context-specific multi-omics data and genetic association studies (such as eQTLs, hQTLs and GWAS SNPs) revealed a series of novel discoveries, including the detection of super long-range QTLs (>900kb), single-cell specific genetic associations, spatial 3D hubs of orchestrated regulatory activities around SNPs, and improved prioritization of causal SNPs. Beyond 1D genomic analysis and traditional association studies, our new predictive and analytical framework opens up a new paradigm to decipher the molecular mechanisms of SNPs associated with complex human diseases from the systems-level view of high-resolution 3D genomes across diverse cellular contexts.

**Novel Ensemble Feature Selection Approach and Application in Repertoire Sequencing Data**

♦*Tao He*1, *Li Zhang*2

1San Francisco State University 2University of California, San Francisco

The T and B cell repertoire make up the adaptive immune system and is mainly generated through somatic V(D)J gene recombination. Thus, the VJ gene usage may be a potential prognostic or predictive biomarker. However, analysis of the adaptive immune system is challenging due to the heterogeneity of the clonotypes that make up the repertoire. To address the heterogeneity of the T and B cell repertoire, we proposed a novel ensemble feature selection approach and customized statistical learning algorithm focusing on the VJ gene usage. We applied the proposed approach to T cell receptor sequences from recovered COVID-19 patients and healthy donors, as well as a group of lung cancer patients who received immunotherapy. Our approach identified distinct VJ genes used in the COVID-19 recovered patients comparing to the healthy donors and the VJ genes associated with the clinical response in the lung cancer patients. Simulation studies show that the ensemble feature selection approach outperformed other state-of-the-art feature selection methods based on both efficiency and accuracy. It consistently yielded higher stability and sensitivity with lower false discovery rates. When integrated with different classification methods, the ensemble feature selection approach had the best prediction accuracy. In conclusion, the proposed novel approach and the integration procedure is an effective feature selection technique to aid in correctly classifying different subtypes to better understand the signatures in the adaptive immune response associated with disease or the treatment in order to improve treatment strategies.

**Empirical Likelihood Based Efficient Semiparametric Inference for Longitudinal Data with Application to GAW 18 Data**

*Yishan Cui*1, ♦*Honglang Wang*1

1Indiana University-Purdue University Indianapolis

We enhance the semiparametric profile estimator for analyzing longitudinal data, specifically addressing the challenge of within-subject correlation. Our refined estimator significantly boosts the efficiency compared to the traditional local kernel smoothing estimator, which operates under the assumption of an independent correlation structure. By integrating a nonparametric operator-regularized approach for estimating the covariance function, we demonstrate that the updated semiparametric estimator still reaches the semiparametric efficiency bound. Additionally, we introduce an Empirical Likelihood (EL)-based method to enable efficient inference. Our evaluation, comprising both simulation studies and an empirical analysis using the Genetic Analysis Workshop 18 dataset, confirms the superior performance of our proposed techniques in a variety of sample sizes and real-world scenarios.

**Session 24CHI55: Novel Applications of Advanced Statistical Learning Methods to Biomedical Data**

**Systems Approach for Congruence and Selection of Cancer Models Towards Precision Medicine**

♦*Jian Zou*1

Cancer models are instrumental as a substitute for human studies and to expedite basic, translational, and clinical cancer research. For a given cancer type, a wide selection of models, such as cell lines, patient-derived xenografts, organoids and genetically modified murine models, are often available to researchers. However, how to quantify their congruence to human tumors and to select the most appropriate cancer model is a largely unsolved issue. Here, we present Congruence Analysis and Selection of CAncer Models (CASCAM), a statistical and machine learning framework for authenticating and selecting the most representative cancer models in a pathway-specific manner using transcriptomic data. CASCAM provides harmonization between human tumor and cancer model omics data, systematic congruence quantification, and pathway-based topological visualization to determine the most appropriate cancer model selection. The systems approach is presented using invasive lobular breast carcinoma (ILC) subtype and suggesting CAMA1 followed by UACC3133 as the most representative cell lines for ILC research. Two additional case studies for triple negative breast cancer (TNBC) and patient-derived xenograft/organoid (PDX/PDO) are further investigated. CASCAM is generalizable to any cancer subtype and will authenticate cancer models for faithful non-human preclinical research towards precision medicine.

**Integrative Mulit-omics Data to Define Three Major Subtypes of Neuroblastoma**

*Jinhua Fan1,* ♦*Yupeng Cun.*

*1Chongqing*

The clinical course of neuroblastoma (NB) varies greatly, ranging from spontaneous regression to malignant progression accompanied by metastatic spread. This indicates that NB comprises different subtypes. We collected gene expression data from 498 patients of NB from the GEO dataset GSE49710. ssGSEA algorithm was used to calculated 28 immune cell feature scores based top rank gene signatures selected by our stSVM . Subsequently, we performed unsupervised clustering analysis using the Similarity Network Fusion (SNF) algorithm on the ssGSEA-calculated immune cell scores and the merged mRNA matrix incorporating the PPI network, for NB patient stratification. The method ensures that the infiltration of immune cells is classified more stably across different datasets and preserves the original matrix information of the samples. Among them, the neurogenic subtype had best survival prognosis, followed by the MES subtype, while the MYCN subtype had the worst prognosis. The MES subtype exhibited high immune cell infiltration scores, low tumor purity, and high expression of mesenchymal transcription factors. Additionally, single-cell data from 16 neuroblastoma patients were collected, and these 16 patients were classified into neurogenic, MES, and MYCN subtypes based similarity measurement of gene signatures, and provide the immune microenvironment characteristics of three new defined neuroblastoma subtypes.

**Bregman Divergence-based Transfer Learning with Application to Survival Model Updating and Data Integration**

♦*Kevin He*1 *and Di Wang*1.

1Department of Biostatistics, University of Michigan.

Prediction of time-to-event data often suffers from rare event rates, small sample sizes, high dimensionality and low signal-to-noise ratios. Incorporating published prediction models from large-scale studies is expected to improve the performance of prognosis prediction on internal small-sized time-to-event data. To account for challenges including heterogeneity, data sharing, and privacy constraints, I will introduce a transfer learning procedure based on Bregman divergence, which measures the discrepancy between the published models and the internal dataset. The proposed procedure is computationally efficient for high-dimensional problems and can be easily implemented with various machine learning methods. Asymptotic properties and simulation results show the advantage of the proposed method compared with those solely based on the internal data. We apply the proposed method to improve prediction performance on a kidney transplant dataset from a local hospital by integrating this small-scale dataset with published survival models obtained from the national transplant registry.

**Bias Correction Models for Electronic Health Records Data in the Presence of Non-random Sampling**

♦*Judy Zhong*1

1NYU School of Medicine

Electronic health records (EHRs) contain rich clinical information for millions of patients and are increasingly used for public health research. However, non-random inclusion of subjects in EHRs can result in selection bias, with factors such as demographics, socioeconomic status, healthcare referral patterns, and underlying health status playing a role. While this issue has been well-documented, little work has been done to develop or apply bias-correction methods, often due to the fact that most of these factors are unavailable in EHRs. To address this gap, we propose a series of Heckman type bias correction methods by incorporating social determinants of health selection covariates to model the EHR non-random sampling probability. Through simulations under various settings, we demonstrate the effectiveness of our proposed method in correcting biases in both the association coefficient and the outcome mean. Our method augments the utility of EHRs for public health inferences, as we show by estimating the prevalence of cardiovascular disease and its correlation with risk factors in the New York City network of EHRs.

**Session 24CHI116: Statistical Innovations in Health Outcomes: Risk Predictions, Treatment Effects, Advanced Regression Techniques and Deep Learning Methodologies**

**Joint Modeling of Interval Censored Adenoma Data and Informative Screening to Predict Risk of Advanced Adenoma**

♦*Yipeng Wei1, May AlHusseini1, Qing Pan1, Hormudz katki2, Rajeshwari Sundaram3*

1The George Washington University， 2National Cancer Institute， 3National Institute of Child Health and Human Development

Recurrent interval-censored (panel count) data is one of the common forms of screening data in epidemiology studies. In the context of colorectal cancer screening, our work is focused on the prediction of the probability of advanced adenoma and risk factors assessment for colorectal cancer. The approach involves non-stationary Poisson process for adenoma and informative screening events, with a semi-parametric Cox model correlated by a latent frailty variable. The study employs the non-parametric Turnbull algorithm for estimating the cumulative distribution function of the intensity function and utilizes the borrow-strength method for estimating subject-specific latent frailty. Marginal prediction model and frailty prediction model are proposed based on the availability of patient’s screening history.

**Enhanced Polytomous Logistic Regression Model for Multicategory Outcomes**

♦*Sheng Fu*1, *Mark P Purdue*1, *Han Zhang*1, *Jing Qin*2, *Lei Song*1, *Kai Yu*1

1National Cancer Institute，2National Institute of Allergy and Infectious Diseases

There is a growing interest in enhancing the efficiency of statistical inference by incorporating aggregated summary information from various sources. However, the integration of summary data and individual-level data for multicategory outcomes poses challenges due to the sophisticated nature of datasets. In this study, we introduce a powerful procedure tailored for multicategory data integration, called PolyGIM. PolyGIM combines polytomous logistic regression (PLR) for modeling individual-level data with summary data generated from two types of logistic regression models. We investigate the theoretical properties of PolyGIM and showcase its advantages through extensive simulation studies. Applying PolyGIM to eight genome-wide association studies within the non-Hodgkin lymphoma (NHL) consortium, we assess the impact of a polygenic risk score associated with lymphoid malignancies on the risks of four NHL subtypes. Our results highlight PolyGIM as a valuable tool for harmonizing data from diverse sources, enabling a more comprehensive evaluation of disease subtype heterogeneity.

**CeCNN: Copula-enhanced Convolutional Neural Networks in Joint Prediction of Refraction Error and Axial Length Based on Ultra-widefield Fundus Images**

♦Chong Zhong1, Yang Li2, Danjuan Yang2, Meiyan Li2, Xingtao Zhou2, Catherine Liu1 and Alan Welsh3

1The Hong Kong Polytechnic University， 2Fudan University， 3Australian National University

Ultra-widefield (UWF) fundus images are replacing traditional fundus images in screening, detection, prediction, and treatment of complications related to myopia because their much broader visual range is advantageous for highly myopic eyes. Spherical equivalent (SE) is extensively used as the main myopia outcome measure, and axial length (AL) has drawn increasing interest as an important ocular component for assessing myopia. Cutting-edge studies show that SE and AL are strongly correlated. Using the joint information from SE and AL is potentially better than using either separately. In the deep learning community, though there is research on multiple-response tasks with a 3D image biomarker, dependence among responses is only sporadically taken into consideration. Inspired by the spirit that information extracted from the data by statistical methods can improve the prediction accuracy of deep learning models, we formulate a class of multivariate response regression models with a higher-order tensor biomarker, for the bivariate tasks of regression-classification and regression-regression. Specifically, we propose a copula-enhanced convolutional neural network (CeCNN) framework that incorporates the dependence between responses through a Gaussian copula (with parameters estimated from a warm-up CNN) and uses the induced copula-likelihood loss with the backbone CNNs. We establish the statistical framework and algorithms for the aforementioned two bivariate tasks. We show that the CeCNN has better prediction accuracy after adding the dependency information to the backbone models. The modeling and the proposed CeCNN algorithm are applicable beyond the UWF scenario and can be effective with other backbones beyond ResNet and LeNet.

**Testing Heterogeneous Treatment Effect with Quantile Regression under Covariate-Adaptive Randomization**

♦*Yang Liu*1, *Lucy Xia*2, *Feifang Hu*3 1Renmin University of China，2Hong Kong University of Science and Technology， 3The George Washington University

In economic studies and clinical trials, it is prevalent to observe heterogeneous treatment effects that vary depending on the relative locations of units in the distribution of responses. In this study, we propose using quantile regression to estimate and conduct inference for the conditional quantile treatment effects (cQTEs) in covariate-adaptive randomized experiments. First, we present sufficient conditions for consistently estimating the cQTEs, concerning the bias due to omitting important covariates in the inference stage. Second, we derive the weak convergence of the quantile regression process and develop a covariate-adaptive randomized bootstrap (CAR-BS) for standard error estimation. Our theoretical results indicate that the Wald test adjusted by CAR-BS is valid for a large class of covariate-adaptive randomization procedures at different quantiles, regardless of the choice of covariates used in inference. We perform extensive numerical and empirical studies to demonstrate advantages of the new method in various settings.

**Session 24CHI76: Recent Advances in Statistical Methods for Complex Imaging Data**

**Density-on-Density Regression**

♦*Yi Zhao*1, *Abhirup Datta*2, *Bohao Tang*2, *Vadim Zipunnikov*3, *Brian Caffo*3

1Indiana University School of Medicine， 2Johns Hopkins University Bloomberg School of Public Health， 3Johns Hopkins University Bloomberg School of Public Health

In this study, a density-on-density regression model is introduced, where the association between densities is elucidated via a warping function. The proposed model has the advantage of a being straightforward demonstration of how one density transforms into another. Using the Riemannian representation of density functions, which is the square root function (or half density), the model is defined in the correspondingly constructed Riemannian manifold. To estimate the warping function, it is proposed to minimize the average Hellinger distance, which is equivalent to minimizing the average Fisher-Rao distance between densities. An optimization algorithm is introduced by estimating the smooth monotone transformation of the warping function. Asymptotic properties of the proposed estimator are discussed. Simulation studies demonstrate the superior performance of the proposed approach over competing approaches in predicting outcome density functions. Applying to a proteomic-imaging study from the Alzheimer’s Disease Neuroimaging Initiative, the proposed approach illustrates the connection between the distribution of protein abundance in the cerebrospinal fluid and the distribution of brain regional volume. Discrepancies among cognitive normal subjects, patients with mild cognitive impairment, and Alzheimer’s disease (AD) are identified and the findings are in line with existing knowledge about AD.

**Functional Support Vector Machine with Applications to Brain Imaging Data**

*Shanghong Xie*1, ♦*Todd Ogden*2

1Southwestern University of Finance and Economics， 2Columbia University

Linear and generalized linear scalar-on-function modeling have been commonly used to understand the relationship between a scalar response variable (e.g., continuous, or binary outcome) and functional or image-valued predictors. Such techniques are sensitive to model misspecification when the relationship between the response variable and the functional predictors is complex. On the other hand, support vector machines (SVMs) are among the most robust prediction models but do not take account of the high correlations between repeated measurements and cannot be used for irregular data. A novel method is proposed to integrate functional principal component analysis (FPCA) with SVM techniques for classification and regression to account for the continuous nature of functional data and the nonlinear relationship between the scalar response variable and the functional predictors. The performance of the method is demonstrated through extensive simulation experiments and through the problem of classification of alcoholics using electroencephalography (EEG) signals.

**Covariate-guided Bayesian Mixture of Spline Experts for the Analysis fNIRS**

*Haoyi Fu1, Lu Tang2, Ori Rosen3, Alison Hipwell2, Theodore Hupperty2,* ♦*Robert Krafty4*

*1*Novartis， *2*University of Pittsburgh， *3*University of Texas El Pas*o 4*Emory University

With rapid development of techniques to measure brain activity and structure, statistical methods for analyzing modern brain-imaging data play an important role in the advancement of science. Imaging data such as functional near-infrared spectroscopy (fNIRS) that measure brain function are usually multivariate high-density longitudinal data and are heterogeneous across both imaging sources and subjects, which lead to various statistical and computational challenges. In this article, we propose a group-based method to cluster a collection of multivariate high-density longitudinal data via a Bayesian mixture of smoothing splines. Our method assumes each multivariate high-density longitudinal trajectory is a mixture of multiple components with different mixing weights. Time-independent covariates are assumed to be associated with the mixture components and are incorporated via logistic weights of a mixture-of-experts model. We formulate this approach under a fully Bayesian framework using Gibbs sampling where the number of components is selected based on a deviance information criterion. The proposed method is compared to existing methods via simulation studies and is applied to a study on fNIRS, which aims to understand infant emotional reactivity and recovery from stress. The results reveal distinct patterns of brain activity, as well as associations between these patterns and selected covariates.

**Motion-Invariant Variational Auto-Encoding of Brain Structural Connectomes**

*Yizi Zhang*, *Meimei Liu*, ♦*Zhengwu Zhang*, *David Dunson*

Mapping of human brain structural connectomes via diffusion MRI offers a unique opportunity to understand brain structural connectivity and relate it to various human traits, such as cognition. However, the presence of motion artifacts during image acquisition can compromise the accuracy of connectome reconstructions and subsequent inference results. We develop a generative model to learn low-dimensional representations of structural connectomes that are invariant to motion artifacts, so that we can link brain networks and human traits more accurately, and generate motion-adjusted connectomes. We applied the proposed model to data from the Adolescent Brain Cognitive Development (ABCD) study and the Human Connectome Project (HCP) to investigate how our motion-invariant connectomes facilitate understanding of the brain network and its relationship with cognition. Empirical results demonstrate that the proposed motion-invariant variational auto-encoder (inv-VAE) outperforms its competitors in various aspects. In particular, motion-adjusted structural connectomes are more strongly associated with a wide array of cognition-related traits than other approaches without motion adjustment.

**Session 24CHI95: Recent Statistical Methodology Advances in Genomics and Bioinformatics**

**CAT: A Conditional Association Test for Microbiome Data Using a Leave-out Approach**

♦*Yushu Shi1, Liangliang Zhang2, Kim-Anh Do3, Robert Jenq3, Christine Peterson3*

1Weill Cornell Medicine， 2Case Western Reserve University， 3University of Texas MD Anderson Cancer Center

In microbiome analysis, researchers often seek to identify taxonomic features associated with an outcome of interest. However, microbiome features are intercorrelated and linked by phylogenetic relationships, making it challenging to assess the association between an individual feature and an outcome. Researchers have developed global tests for the association of microbiome profiles with outcomes using beta diversity metrics which offer robustness to extreme values and can incorporate information on the phylogenetic tree structure. Despite the popularity of global association testing, most existing methods for follow-up testing of individual features only consider the marginal effect and do not provide relevant information for the design of microbiome interventions. This paper proposes a novel conditional association test, CAT, which can account for other features and phylogenetic relatedness when testing the association between a feature and an outcome. CAT adopts a leave-out method, measuring the importance of a feature in predicting the outcome by removing that feature from the data and quantifying how much the association with the outcome is weakened through the change in the coefficient of determination. By leveraging global tests including PERMANOVA and MiRKAT-based methods, CAT allows association testing for continuous, binary, categorical, count, survival, and correlated outcomes. Our simulation and real data application results illustrate the potential of CAT to inform the design of microbiome interventions aimed at improving clinical outcomes.

**spVC for the Detection and Interpretation of Spatial gene Expression Variation**

*Shan Yu*1, ♦*Wei Vivian Li*2.

1University of Virginia， 2University of California, Riverside.

Recent advances in spatially resolved transcriptomics technologies have opened up new avenues for understanding gene expression heterogeneity in spatial contexts, and an important task in spatial transcriptomics analysis is the identification of spatially variable genes (SVGs). While various computational methods exist for SVG detection, most focus solely on statistical significance and have limitations in capturing continuous expression patterns across spatial domains and incorporating cell/spot-level covariates. To address these challenges, we introduce spVC, a novel statistical method to detect and interpret SVGs based on a generalized Poisson model. spVC integrates constant and spatially varying effects of cell/spot-level covariates, enabling comprehensive exploration of gene expression variability and enhancing interpretability. It provides a convenient tool to identify potential factors that contribute to gene expression variability, including spatial locations and other cell/spot-level covariates such as cell types or tissue layers. It offers estimation and statistical inference tools for both constant and spatially varying coefficients, allowing for the selection of different types of SVGs. In summary, spVC is a versatile tool for the identification, interpretation, and comprehension of gene expression variation in spatial transcriptomics data.

**Robust Mendelian Randomization Analysis by Automatically Selecting Valid Genetic Instruments with Applications to Identify Plasma Protein Biomarkers for Alzheimer’s Disease**

*Minhao Yao1, Gary Miller2, Badri Vardarajan2, Andrea Baccarelli3*, ♦*Zhonghua Liu4*

1University of Hong Kong， 2Columbia University， 3Harvard University， 4Rutgers， 5Columbia University

Mendelian randomization (MR) uses genetic variants as instrumental variables (IVs) to infer the causal effect of a modifiable exposure on the outcome of interest by removing unmeasured confounding bias. However, some genetic variants might be invalid IVs due to violations of core IV assumptions. MR analysis with invalid IVs might lead to biased causal effect estimate and misleading scientific conclusions. To address this challenge, we propose a novel MR method that first Selects valid genetic IVs and then performs Post-selection Inference (MR-SPI) based on two-sample genome-wide summary statistics. We analyze 912 plasma proteins using the large-scale UK Biobank proteomics data in 54,306 participants and identify 7 proteins (TREM2, PILRB, PILRA, EPHA1, CD33, RET, CD55) significantly associated with the risk of Alzheimer’s disease. We employ AlphaFold2 to predict the 3D structural alterations of these 7 proteins due to missense genetic variations, providing new insights into their biological functions in disease etiology.

**Statistical Tools for the Analysis of Differential m6A Methylation**

♦*Zhenxing Guo*

CUHK-Shenzhen

RNA methylation has emerged recently as an active research domain to study post-transcriptional alteration in gene expression regulation. One of the fundamental questions in RNA methylation data analysis is to identify the Differentially Methylated Regions (DMRs), by contrasting cases and controls. In this talk, we will present TRESS which utilizes the hierarchical negative binomial model to examine changes in m6A methylation across various conditions, which is followed by the work of thoroughly benchmarking eight existing methods for DMR calling using both synthetic and real data. Then based upon the benchmark performance of available methods, we further present a statistical power assessment tool, MAGPIE, for power calculation and experimental design in epitranscriptome studies using m6A sequencing data.

**Session 24CHI86: Recent Development of Joint Modeling of Longitudinal and Survival Data for Population Health Research**

**On the Time-varying Predictive Performance of Longitudinal Biomarkers and Risk Prediction Models**

*Jing Zhang*1, *Jing Ning*2, *Xuelin Huang*2, ♦*Ruosha Li*1.

1The University of Texas Health Science Center at Houston, 2The University of Texas MD Anderson Cancer Center

In many biomedical studies, participants are monitored at periodic visits until the occurrence of the failure event. Biomarkers are often measured repeatedly during these visits, and such measurements can facilitate updated disease prediction. In this work, we propose a two-dimensional incident dynamic area under curve (AUC), to capture the variability due to both the biomarker assessment time and the prediction time to comprehensively quantify the predictive performance of a longitudinal biomarker. We propose a pseudo partial-likelihood to achieve consistent estimation of the AUC under two realistic scenarios of visit schedules. Variance estimation methods are designed to facilitate inferential procedures. We examine the finite-sample performance of our method through extensive simulations. The methods are also extended to estimate the two-dimensional AUC based on an estimated dynamic prediction model that combines multiple longitudinal biomarkers.

**Improving Estimation Efficiency for Survival Data Analysis by Integrating a Coarsened Time-to-Event Outcome from an External Study**

*♦Chixiang Chen1, Daxuan Deng2, Ming Wang3.*

1University of Maryland, School of Medicine, 2Penn State university, 3Case Western Reserve University

In the era of big data, increasing availability of data makes combining different data sources to obtain more accurate estimations a popular topic. However, the development of data integration is often hindered by the heterogeneity in data forms across studies. In this talk, we focus on a case in survival analysis where we have primary study data with a continuous time-to-event outcome and complete covariate measurements, while the data from an external study contain an outcome observed at regular intervals and only a subset of covariates is measured. To incorporate external information while accounting for the different data forms, we posit working models and obtain informative weights by empirical likelihood, which will be used to construct a weighted estimator in the main analysis. We have established the theory demonstrating that the new estimator has higher estimation efficiency compared to the conventional ones, and this advantage is robust to working model misspecification, as confirmed in our simulation studies. To assess its utility, we apply our method to accommodate data from the National Alzheimer’s Coordinating Center to improve the analysis of the Alzheimer’s Disease Neuroimaging Initiative phase 1 study.

**The Backward Joint Model of Multivariate Longitudinal Trajectories and Multivariate Survival Outcomes with Application in Dynamic Prediction Problems**

*♦Liang Li*1*.*

1The University of Texas MD Anderson Cancer Center

Joint modeling is an important approach to dynamic prediction of clinical outcomes using longitudinally measured predictors, such as biomarkers. We consider the situation where the predictors include baseline covariates and the longitudinal trajectories of many correlated biomarkers, measured asynchronously at irregularly spaced time points. The outcomes of predictive interest include both the terminal clinical event with or without competing risks, and the future longitudinal biomarker trajectories if the terminal or competing risk events do not occur. We propose a novel backward joint model (BJM) to solve this problem. The BJM can be flexibly specified to optimize the prediction accuracy. Its likelihood-based estimation algorithm is robust, fast, and stable regardless of the dimension of longitudinal biomarkers. We illustrate the BJM methodology with simulations and a real dataset from the African American Study of Kidney Disease and Hypertension.

**Joint Models of Longitudinal and Survival Data with Censoring and Outliers**

♦*Lang Wu1.*

*1*University of British.

In a survival model with a time-dependent covariate, the covariate may be left censored due to a lower detection limit and its observed values may contain outliers. Motivated from an HIV vaccine study, we propose a robust method for joint models of longitudinal and survival data, where the outliers in longitudinal data are addressed using a multivariate t-distribution for b-outliers and using an M-estimator for e-outliers. We also propose a computationally efficient method for approximate likelihood inference. The proposed method is evaluated by simulation studies. Based on the proposed models and method, we analyze the HIV vaccine data and find a strong association between longitudinal biomarkers and the risk of HIV infection.

**Session 24CHI78: Recent Advances in Statistical Methods with Applications in Biomedical Research**

**An Innovative Centile Chart Method for Utilizing Natural History Data in Rare Disease Clinical Development**

♦*Tianying Wang1, Wenfei Zhang2 and Ying Wei3.*

1Colorado State University, 2Sarepta Therapeutics， 3Columbia University

Utilizing natural history data as external control plays an important role in the clinical development of rare diseases, since placebo groups in double-blind randomization trials may not be available due to ethical reasons and low disease prevalence. This article proposed an innovative approach for utilizing natural history data to support rare disease clinical development by constructing reference centile charts. Due to the deterioration nature of certain rare diseases, the distributions of clinical endpoints can be age-dependent and have an absorbing state of zero, which can result in censored natural history data. Existing methods of reference centile charts can not be directly used in the censored natural history data. Therefore, we propose a new calibrated zero-inflated kernel quantile (ZIKQ) estimation to construct reference centile charts from censored natural history data. Using the application to Duchenne Muscular Dystrophy drug development, we demonstrate that the reference centile charts using the ZIKQ method can be implemented to evaluate treatment efficacy and facilitate a more targeted patient enrollment in rare disease clinical development.

**A Partially Functional Linear Regression Framework for Integrating Genetic, Imaging, and Clinical Data**

♦*Ting Li1 and Hongtu Zhu2.*

1UShanghai University of Finance and Economics， 2University of North Carolina at Chapel Hill

This paper is motivated by the joint analysis of genetic, imaging, and clinical (GIC) data collected in the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study. We propose a partially functional linear regression (PFLR) framework to map high-dimensional GIC-related pathways for Alzheimer’s disease (AD). We develop a joint model selection and estimation procedure by embedding imaging data in the reproducing kernel Hilbert space and imposing the L0 penalty for the coefficients of genetic variables. We apply the proposed method to the ADNI dataset to identify important features from tens of thousands of genetic polymorphisms (reduced from millions using a pre- processing step) and study the effects of a certain set of informative genetic variants and the baseline hippocampus surface on 13 future cognitive scores. We also explore the shared and distinct heritability patterns of these cognitive scores. Analysis results suggest that both the hippocampal and genetic data have heterogeneous effects on different scores, with the trend that the value of both hippocampi are negatively associated with the severity of cognition deficits. Polygenic effects are observed for all the thirteen cognitive scores. The well-known APOE4 genotype only explains a small part of the cognitive function. Shared genetic etiology exists; however, greater genetic heterogeneity exists within disease classifications after accounting for the baseline diagnosis status. These analyses are useful in further investigation of functional mechanisms for AD progression.

**Efficient Designs and Analysis of Two-Phase Studies with Longitudinal Binary Data**

♦*Ran Tao1.*

1Vanderbilt University Medical Center

Researchers interested in understanding the relationship between a readily available longitudinal binary outcome and a novel biomarker exposure can be confronted with ascertainment costs that limit sample size. In such settings, two-phase studies can be cost-effective solutions that allow researchers to target informative individuals for exposure ascertainment and increase estimation precision for time-varying and/or time-fixed exposure coefficients. In this paper, we introduce a novel class of residual-dependent sampling (RDS) designs that select informative individuals using data available on the longitudinal outcome and inexpensive covariates. Together with the RDS designs, we propose a semiparametric analysis approach that efficiently uses all data to estimate the parameters. We describe a numerically stable and computationally efficient EM algorithm to maximize the semiparametric likelihood. We examine the finite sample operating characteristics of the proposed approaches through extensive simulation studies, and compare the efficiency of our designs and analysis approach with existing ones. We illustrate the usefulness of the proposed RDS designs and analysis method in practice by studying the association between a genetic marker and poor lung function among patients enrolled in the Lung Health Study.

**CompDA: Defining and Finding a New Type of Health-Microbiome Associations**

♦*Siyuan Ma1*.

1Vanderbilt University Medical Center.

A major task of microbiome epidemiology is association analysis, where the goal is to identify microbial features related to host health. This is commonly performed by the differential abundance (DA) analysis, which, by design, examines each microbe as isolated from the rest of the microbiome. This does not properly account for the microbiome’s compositional nature or microbe-microbe ecological interactions, and can lead to confounded findings, i.e., microbes that only appear to associate with health through their confounding association with health-related, biologically informative microbes. To remedy these issues, we present Compositional Differential Abundance (CompDA) analysis, a novel approach for health-microbiome association. CompDA provides a novel approach to identify health-related microbes by examining the microbiome holistically, which a) accounts for the data’s compositionality and ecological interactions, and b) has clear interpretations corresponding to host health as affected by microbiome-based interventions. CompDA prioritizes health-related microbes and controls false discoveries by implementing recent advances from high-dimensional statistics, and can be flexibly adapted to many common tasks in modern microbiome epidemiology, including enhancing microbiome-based machine learning by providing rigorous p-values to prioritize important features. We validate the performance of CompDA, and compare against canonical microbiome association methods including DA with extensive, real-data-informed simulation studies. Lastly, we report novel and consistent findings of CompDA in application studies, including a) recently reported microbial signatures of colorectal cancer from cross-study machine learning, and b) well-established microbial associations of early onset Crohn’s disease in a pediatric cohort.

**Session 24CHI92: Recent Developments in Statistical Machine Learning**

**Conditional Modeling of Panel Count Data with Partly Interval-censored Failure Event**

♦*Xiangbin Hu*1, *Wen Su*2, *Zhisheng Ye*3 and *Xingqiu Zhao*1*.*

1The Hong Kong Polytechnic University， 2City University of Hong Kong， 3National University of Singapore

In longitudinal follow-up studies, panel count data arise from discrete observations on recurrent events. We investigate a more general situation where a partly interval-censored failure event is informative to recurrent events. The existing methods for the informative failure event are based on the latent variable model, which provides indirect interpretation for the effect of failure event. To solve this problem, we propose a failure time-dependent proportional mean model with panel count data through an unspecified link function. For estimation of model parameters, we consider a conditional expectation of least squares function to overcome the challenges from partly interval-censoring, and develop a two-stage estimation procedure by treating the distribution function of the failure time as a functional nuisance parameter and using the B-spline functions to approximate unknown baseline mean and link functions. Furthermore, we derive the overall convergence rate of the proposed estimators and establish the asymptotic normality of finite-dimensional estimator and functionals of infinite-dimensional estimator. The proposed estimation procedure is evaluated by extensive simulation studies, in which the finite-sample performances coincide with the theoretical results. We further illustrate our method with a longitudinal healthy longevity study and draw some insightful conclusions.

**Variable Selection with Deep Differential Neural Networks**

♦*Yu Chen*1, *Jian Huang*1, *Guohao Shen*1 and *Xingqiu Zhao*1*.*

1The Hong Kong Polytechnic University

This paper explores the topic of variable selection using deep differentiable neural networks. We consider the loss function with a general form, which includes least square loss, logistic loss and other loss that meet the loss assumption. The proposed method is demonstrated through derivative-induced sparsity, which avoids relying on explicit model assumptions. Leveraging the advantageous mathematical properties of deep differential neural networks, we establish the convergence rate of the nonparametric estimator. Additionally, we derive a corresponding derivative bound, ensuring selection consistency. Through a comprehensive simulation study, we demonstrate the effectiveness of the proposed method in practical scenarios. Finally, we apply the proposed approach to real-world datasets that served as the motivation for this study.

**Deep Generalized Accelerated Hazards Model with Interval-censored Data**

♦*Qiang Wu*1, *Mingyue Du*1 *and Xingqiu Zhao*1

1The Hong Kong Polytechnic University 2Jilin University

For the analysis of interval-censored data, we introduce a more comprehensive model called the generalized accelerated hazards model. This model aims to facilitate a thorough analysis of the relationship between various risk factors and the hazard of the failure time. We propose a maximum likelihood estimation procedure that combines the use of fully connected neural networks and monotonic splines. We leverage fully connected neural networks to approximate the nonparametric effects, allowing for flexible and adaptive modeling of complex relationships. In addition, we employ monotonic splines to approximate the baseline cumulative hazard function, ensuring that the estimated function maintains a monotonic behavior. Furthermore, under certain regular conditions, we obtain the rate of convergence of the overall parameter estimate and establish the asymptotic normality of the parameter estimates. To evaluate the performance of our proposed approach, we conducted a simulation study to assess its finite sample properties. Furthermore, as a real-world application, our proposed method is applied to the Atherosclerosis Risk in Communities (ARIC) study.

**Strategic Statistical Learning for Some Applications**

♦*Xiaodong Yan*

Nonlinear expectation is an original research direction developed in China, which is increasingly important for scientific research in various fields, especially with the rise of big data and artificial intelligence, providing a stronger impetus for innovative theory and application research of nonlinear expectations. Recently, our team has developed the "Strategy Limit Theory" based on the simplest reinforcement learning model - the multi-arm bandit model, which is a significant contribution in cross-research between nonlinear probability theory and some statistical learning. It has changed the traditional statistical research paradigm. Subsequent related research has made significant contributions in interpretable and responsible statistical theories and methods such as big data sampling, experimental design, transfer learning, online learning, meta learning and so on.