WNAR/IMS 2022 Abstracts

Last updated 2022-05-16

IMS Invited Sessions

IMS Invited 1

New frontiers in nonparametric learning, sparse learning, and deep learning

Organizer & Chair: Xinyi Li, School of Mathematical and Statistical Sciences, Clemson University

Dimension reduction of longitudinal microbiome data by tensor functional SVD

Rungang Han (Duke University), Pixu Shi (Duke University), Anru Zhang (Duke University)*

The reduction of sequencing cost has prompted more microbiome studies with longitudinal measurements of bacterial abundance. Longitudinal microbiome data can often be formatted into a high-dimensional order-3 tensor with three modes representing the subject, time, and bacteria respectively. Since the time of measurement for different subjects can be highly variable, the values of such order-3 tensor are typically not well-aligned, making it challenging to analyze the trajectory of bacterial abundance over time and identify key bacteria associated with time or clinical phenotypes. In this paper, we propose a new tensor functional SVD method that performs dimension reduction to assist the analysis of high-dimensional¬†longitudinal microbiome data. The new method can extract the key components in the trajectories of bacterial abundance, identify representative bacterial taxa for these key trajectories, and group subjects based on the change of bacteria abundance over time. The new method is also flexible to handle microbiome measurements at irregular time points for different subjects.

Interval Privacy: A New Framework for Privacy-Preserving Data Collection

Jie Ding (University of Minnesota), Bangjun Ding (East China Normal University)*

The emerging public awareness and government regulations of data privacy motivate new paradigms of collecting and analyzing data transparent and acceptable to data owners. This talk will introduce a new concept of privacy and related data formats, mechanisms, and theories for statistically privatizing data during data collection. The new privacy mechanisms will record each data value as a random interval (or, more generally, a range) containing it. Such mechanisms can be easily deployed through survey-based data collection interfaces, e.g., by asking a respondent whether its data value is within a randomly generated range. Using narrowed range to convey information is complementary to the existing paradigm of randomized response. Also, the proposed mechanisms can generate progressively refined information at the discretion of individuals, naturally leading to privacy-adaptive data collection. This talk will demonstrate unique perspectives brought by Interval Privacy for human-centric data privacy, where individuals enjoy a perceptible, transparent, and simple way of sharing sensitive data.

Optimal and Safe Estimation for High-Dimensional Semi-Supervised Learning

Siyi Deng (Cornell University), Yang Ning (Cornell University), Jiwei Zhao (University of Wisconsin-Madison), Heping Zhang (Yale University)*

We consider the estimation problem in high-dimensional semi-supervised learning. The data with the observed outcomes are called labeled, and those without the outcomes are referred to as unlabeled. In this paper, we consider the linear regression problem with such a data structure under the high dimensionality. Our goal is to investigate when and how the unlabeled data can be exploited to improve the estimation of the regression parameters of linear model in light of the fact that such linear models may be misspecified in data analysis. We first establish the minimax lower bound for parameter estimation in the semi-supervised setting. We show that the supervised estimators using the labeled data only cannot attain this lower bound. When the conditional mean function is correctly specified, we

propose an optimal semi-supervised estimator which attains the lower bound and therefore improves the rate of the supervised estimators. To alleviate the strong requirement for this optimal estimator, we further propose a safe semi-supervised estimator.

Large Scale Prediction with Decision Trees

Jason Klusowski (Princeton University)*

Decision trees are one of the most elemental methods for classification problems. They are often deployed in contexts where many explanatory variables are observed and where a high importance is placed on the simplicity and interpretability of the fitted model, such as business and medicine. In this talk, I will show that trees trained with C4.5 methodology are consistent for certain nonparametric classification models, even when the number of explanatory variables grows sub-exponentially with the sample size, under both weak and strong forms of sparsity. These statistical guarantees highlight the crucial role that data dependent splits play in determining the adaptive properties of the tree.

WNAR Invited Sessions

WNAR Invited 1

Semi- and non- parametric statistical methods for omics data

Organizer & Chair: Xinlian Zhang, University of California, San Diego

Distance-based Semiparametric Regression Framework for Between-subject attributes: Applications to High-dimensional Sequences of Microbiome and Wearables

Jinyuan Liu (Department of Family Medicine and Public Health, UC San Diego, San Diego, California, U.S.A.), Xinlian Zhang (Department of Family Medicine and Public Health, UC San Diego, San Diego, California, U.S.A.), Tanya T. Nguyen (Center for Microbiome Innovation, University of California San Diego, San Diego, California, U.S.A.), Dilip V. Jeste (Stein Institute for Research on Aging, UC San Diego, San Diego, California, U.S.A.), Ellen Lee (Stein Institute for Research on Aging, UC San Diego, San Diego, California, U.S.A.), Xin Tu (Department of Family Medicine and Public Health, UC San Diego, San Diego, California, U.S.A.)*

Breakthroughs in high-throughput sequencing and wearable devices are enlightening insights into inherent disease mechanisms, while those high-dimensional data also provoke substantial challenges in statistical analyses and interpretations. As an interesting way to view such data, between-subject attributes that compare two subjects' sequence with dissimilarity metrics are growingly adopted as a dimension-reduction summary. To elucidate complex relationships between such sequences and clinical phenotypes, one needs to discern mixed within- and between-subject variability while simultaneously handling correlations among those pairwise dissimilarities. By extending the generalized linear model (GLM) from within- to between-subject attributes, we present a unified GLM-type regression framework designated for distance metrics. In the burgeoning research fields producing data with astronomical dimensions, this timely solution fills a critical gap by accommodating data of any type or dimension to delineate robust inference. We illustrate its superiority over comparable approaches with simulated and real sequence data from microbiome and wearables, where dimensions are tens of thousands.

Statistical model for recovering the low rank structure of spatial transcriptomics data

Sha Cao (Indiana University)*

Currently, methods specifically designed for spatial transcriptomics (ST) data modeling are lacking. Firstly, for most existing methods, cellular and regional expression profiles are typically analyzed first without the spatial information and only later projected back onto the spatial structure for visual inspection of spatial trend. Secondly, similar to single cell RNA-Seq data, ST based gene expression data is also plagued by dropout events, a phenomenon where genes actually expressed in a given cell or region are incorrectly measured as unexpressed. Thirdly, the gene by sample expression matrix is no longer retainable for many spatial methods. To address these challenges, we present a regularized maximum likelihood estimator to recover the noisy observed expression matrix as an approximately low-rank expression matrix under Poisson distribution, which is also spatially smooth. Our method enables spatial clustering by modeling a low-dimensional representation of the count-based gene expression matrix and encouraging neighboring spots to belong to the same cluster via a spatial smoothness penalty term.

B-scaling: a novel nonparametric data fusion method

Yiwen Liu (University of Arizona), Xiaoxiao Sun (University of Arizona), Wenxuan Zhong (University of Georgia), Bing Li (Pennsylvania State University)*

With the rapid development in science and technology, massive data has been collected from different sources, which leads to a large amount of data with different types and formats, such as the image data and omics data. Each type of the data only captures part of the contained information, and the data has to be integrated or fused to provide a complete understanding of the whole picture. Thus, there is an

urgent call of powerful data fusion method. In this talk, I will introduce a B-scaling method to integrate multisource data. The asymptotic property of the B-scaling method will be discussed to provide theoretical underpinning of the method. The application of the method on epigenetic and biomedical research will be highlighted in the talk.

WNAR Invited 2

Ecological Statistics Early Career Researcher Showcase

Organizer & Chair: Ecological Statistics Early Career Researcher Showcase

Simultaneous estimation of diet composition and covariate modelling: an extension to MUFASA

Holly Steeves (University of Western Ontario), Chris Field (Dalhousie University), Connie Stewart (University of New Brunswick Saint John), Shelley Lang (Department of Fisheries and Oceans), Aaron MacNeil (Dalhousie University)*

Diet compositions of marine predators are often of interest for marine ecologists in trophic structure studies where non-lethal sampling has created a need for non-invasive diet estimation techniques. Methods using fatty acids, including quantitative fatty acid signature analysis (QFASA) and maximum unified fatty acid signature analysis (MUFASA), have been developed to obtain dietary estimates that have previously been difficult to acquire. Unlike QFASA, MUFASA is able to include covariates in the model. With use of a link function, the diet proportions are assumed to be a function of the covariates. This method yields a summary diet for all unique sets of covariates. It also allows for inference on diet estimates between various groups, such as sex, age or environmental factors. Simulations yield accurate summary estimates and inference lead to making the correct decision in all cases run. Finally, these techniques are used to analyze a real life study of grey seals off of Sable Island. Using sex and type of population growth on Sable for each seal, our method yield similar summary estimates to QFASA, and results were in agreement with the beliefs of biologists.

Bayesian state-space models with applications to aquatic acoustic telemetry data

Inesh Munaweera (University of Manitoba), Saman Muthukumarana (University of Manitoba), Darren M. Gillis (University of Manitoba)*

Recent advances in animal tracking technologies such as acoustic telemetry (AT) have enabled researchers to collect enormous amounts of data on animal movement and habitat use over large geographic scales. In this talk, we present two applications of Bayesian state-space modeling (SSM) for AT data to 1) estimate individual-level Walleye moment paths using a systematically deployed two-dimensional array of receivers in Lake Winnipeg 2) estimate the survival and recapture probabilities of Arctic Char in different habitats of Cambridge Bay region using receivers placed at geographical bottlenecks. Here, we show how to handle practical challenges that arise in modeling ecological data obtained using omnidirectional AT systems. Furthermore, we highlight the advantages of the Bayesian SSM approach to each application over classical approaches such as the smoothing techniques to estimate fish moment paths and the Cormack-Jolly-Seber model to estimate survival probabilities. In addition, we will further illustrate how each result of the study provides valuable information in making effective fishery management decisions in the corresponding regions in the future.

N-mixture models for large populations

Matthew Parker (Simon Fraser University), Jiguo Cao (Simon Fraser University), Laura Cowen (University of Victoria), Lloyd Elliott (Simon Fraser University)*

We derive an asymptotic likelihood function for open-population N-mixture models and show that it has favorable computational complexity and accuracy when compared to the traditional likelihood function for large population sizes. Our asymptotic model is validated using simulation studies. We show that our methods perform favorably in both accuracy and precision compared with a multivariate normal approximation which has been previously developed. Our model is applied to estimate the population

size of Ancient Murrelet chicks, comparing against results obtained using the traditional N-mixture likelihood. We provide an open source implementation of our methods in the quickNmix R package.

Spatial Capture-Recapture Without Animal Identity

Paul van Dam-Bates (University of St Andrews), David Borchers (University of St Andrews), Michail Papathomas (University of St Andrews)*

Human observers are quickly becoming replaced by modern technology like camera traps and acoustic recording units that can reliably record animal detections over a long period of time. When animals have a chance to be recorded on multiple detectors, spatial capture-recapture (SCR) can be used to estimate animal density. However, these models generally require that animal identity is known. We reformulate the conventional SCR model as a marked Poisson process such that the counting process for detections through time no longer depends on identity, but the observed mark distributions do. As a result, when identity is latent, the mark observation is a mixture of N latent animal characteristics (e.g. activity centre, sex), where N is the population size. This becomes a generalization of the unmarked SCR model of Chandler and Royle (2013) and allows us to easily add additionally observed covariates to help estimate animal identity. We show through simulation how well the method works and apply it to a camera trap survey of fisher (Pekania pennanti) and an acoustic survey of the Cape Peninsula moss frog (Arthroleptella lightfooti), each with different amounts of mark information.

WNAR Invited 3

Recent Advances in Categorical Data Analysis

Organizer & Chair: Krishna Saha, Central CT State University

High-dimensional fixed effects profiling models: New developments and applications

Danh Nguyen (University of California Irvine), Damla Senturk (UCLA), Jason Estes (Mountain View), Esra Kurum (UC Riverside)*

Profiling analysis aims to evaluate health care providers, such as hospitals, nursing homes, or dialysis facilities etc., with respect to a patient outcome. fixed effects (FE) profiling methods have considered binary outcomes, such as 30-day hospital readmission or mortality. For the unique population of dialysis patients, (1) regular blood tests are required to evaluate the effectiveness of treatment and avoid adverse events, including dialysis inadequacy, imbalance mineral levels, and anemia among others, as well as (2) the need for continuous monitoring/care after transitioning to dialysis. We illustrate the versatility of FE profiling models through several applications in profiling dialysis facilities in the U.S. and recent FE model developments, including (a) time-varying/time-dynamic standardized readmission ratio, (b) profiling for adverse recurrent events, and (c) new insights on operating characteristics such performance of FE model under the low information context/sparse outcome data setting.

Likelihood-based Approach for Testing the Homogeneity of Risk Difference or Risk Ratio in a Multicenter Randomized Clinical Trial

Danh V Nguyen (University of California at Irvine), Yeongjin Gwon (University of Nebraska Medical Center), Mili Roy (University of Calgary), Krishna Saha (Central CT State University)*

Lipsitz et al. (1998) proposed various test procedures for testing the homogeneity of the risk difference in a multicenter randomized clinical trial when the data are sparse. However, in some situations, these test procedures showed serious liberal behaviors. To improve these test procedures, Lui and Kelly (2000) considered three different suggested approaches, but still these test procedures behave very similarly by showing improvement only in power. To overcome these limitations, we develop some likelihood-based test procedures for testing the homogeneity of the risk difference based on the binomial and the beta-binomial models. We also propose to improve the existing test procedures using the correct variance estimators by taking within center correlation into account. Our proposed test procedures are

then compared with the existing test procedures, by Monte Carlo simulations, in terms of size and power. An illustrative application of the proposed test procedures is presented.

New Development in Regression Model for Aggregate Ordinal Outcomes with Missing Categories

Yeongjin Gwon (University of Nebraska Medical Center), Ming-Hui Chen (University of Connecticut), May Mo (Amgen Inc), Tony Jiang (Amgen Inc), Amy Xia (Amgen Inc), Joseph Ibrahim (University of North Carolina)*

The ordinal response variable will inevitably contain unknown response categories because they cannot be directly derived from published data in the literature. In this talk, we propose a statistical methodology to overcome such a common but unresolved issue in the context of network meta-regression for aggregate ordinal outcomes. Specifically, we introduce unobserved latent counts and model these counts within a Bayesian framework. The proposed approach includes several existing models as special cases and also allows us to conduct a proper statistical analysis in the presence of trials with certain missing categories. We then develop an efficient Markov chain Monte Carlo sampling algorithm to carry out Bayesian computation. A variation of the deviance information criterion is used for the assessment of goodness-of-fit under different distributions of the latent counts. A case study demonstrating the usefulness of the proposed methodology is carried out using aggregate ordinal outcome data from 17 clinical trials in treating Crohn's Disease.

Joint Analysis of Binary and Continuous Data via Joint Modelling of Jittered Binary and Continuous Outcomes: A New Approach

Mili Roy (University of Calgary), Dr. Alex de Leon (University of Calgary)*

Important recent work on copula models has opened new directions in the analysis of mixed data, including flexible extensions of widely used mixed data models to accommodate disparate non-Gaussian continuous outcomes and latent variables to jointly analyze non-Gaussian mixed data. However, models that rely on thresholding a latent multivariate meta-distribution, while practically appealing, are not always appropriate in applications that involve count and nominally scaled categorical outcomes. In this talk, we develop a new methodology that entails transforming binary outcomes into continuous by "jittering"– thus reducing the problem to one that involves only continuous variables – and jointly modelling them with continuous outcomes using Gaussian copula mixed models. Although jittering has previously been used in copula modelling of discrete data, it was ostensibly only for alleviating computational difficulties arising from likelihood analysis of discrete data . We explore the finite-sample properties of likelihood-based estimates in simulations and revisit the ethylene glycol mice data for illustration.

WNAR Invited 4

Advanced statistical methods for biomedical data

Organizer & Chair: Esra Kurum, University of California, Riverside

A functional model for studying common trends across trial time in eye tracking experiments

Mingfei Dong (UCLA), Donatello Telesca (UCLA), Catherine Sugar (UCLA), Frederick Shic (University of Washington), James McPartland (Yale University), Damla Senturk (UCLA)*

Eye tracking (ET) experiments commonly record the continuous trajectory of a subject's gaze on a two-dimensional screen throughout repeated presentations of stimuli (referred to as trials). Even though the continuous path of gaze is recorded during each trial, commonly derived outcomes for analysis collapse the data into simple summaries, such as looking times in regions of interest. In order to retain information in trial time, we utilize functional data analysis (FDA) for the first time in literature in the analysis of ET data. More specifically, novel functional outcomes for ET data, referred to as viewing profiles, are introduced that capture the common gazing trends across trial time which are lost in traditional data summaries. Mean and variation of the proposed functional outcomes across subjects are then modeled using functional principal components analysis. Applications to data from a visual

exploration paradigm conducted by the Autism Biomarkers Consortium for Clinical Trials find significant group differences between children diagnosed with autism and their typically developing peers in their consistency of looking at faces early on in trial time.

Multilevel time-varying joint models for hospitalization and survival in patients on dialysis

Esra Kurum (University of California, Riverside)*

Over 720,000 patients with end-stage kidney disease in the U.S. require life-sustaining dialysis treatment. In this population of typically older patients with a high morbidity burden, hospitalization is frequent at about twice per year. Aside from frequent hospitalizations, which is a major source of death risk, overall mortality in these patients is higher than other comparable populations, including Medicare patients with cancer. Thus, understanding patient- and facility-level risk factors that jointly contribute to hospitalizations and mortality is of interest. Towards this objective, we propose a novel methodology to jointly model hospitalization, a binary longitudinal outcome, and survival, based on multilevel data from the United States Renal Data System (USRDS), with repeated observations over time nested in patients and patients nested in dialysis facilities. To accommodate the USRDS data structure, we depart from the joint modeling literature by including multilevel random effects and multilevel covariates. An approximate EM algorithm is developed for estimation and inference where fully exponential Laplace approximations are utilized to address computational challenges.

Discovering high-order interaction with signed iterative random forests

Karl Kumbier (UC San Francisco), Sumanta Basu (Cornell University), Erwin Frise (Lawrence Berkeley National Laboratory), Sue Celniker (Lawrence Berkeley National Laboratory), James B. Brown (Lawrence Berkeley National Laboratory), Bin Yu (UC Berkeley)*

The recent explosion of high-dimensional genomic datasets, paired with powerful supervised learning algorithms, is beginning to elucidate molecular interactions that drive development and function. However, state-of-the-art prediction algorithms are typically black-boxes, offering limited insight into the complex associations they learn. We address this challenge by building on the iterative Random Forest (iRF) algorithm to explicitly map responses as a function of the learned feature interactions. Our method, signed iRF (siRF), describes ``subsets'' of rules that frequently occur on iRF decision paths. We refer to these ``rule subsets'' as signed interactions. Signed interactions share not only the same set of interacting features but also exhibit similar thresholding behavior, and thus describe a stable relationship between interacting features and responses. We describe stable and predictive importance metrics (SPIMs) to rank signed interactions in terms of their stability, predictive accuracy, and strength of interaction. We evaluate our proposed approach in biologically inspired simulations and a case study predicting enhancer activity from TF binding.

Learning from Real-World Data About Combinatorial Treatment Selection for COVID-19

Xinping Cui (University of California, Riverside)*

COVID-19 is an unprecedented global pandemic with a serious negative impact on virtually every part of the world. This paper reports a case study of combinatorial treatment selection for COVID-19 based on real-world data from a large hospital in Southern China. In this observational study, 417 confirmed COVID-19 patients were treated with various combinations of drugs and followed for four weeks after discharge (or until death). Treatment failure is defined as death during hospitalization or recurrence of COVID-19 within four weeks of discharge. Using a virtual multiple matching method to adjust for confounding, we estimate and compare the failure rates of different combinatorial treatments, both in the whole study population and in subpopulations defined by baseline characteristics. Our analysis reveals that treatment effects are substantial and heterogeneous, and that the optimal combinatorial treatment may depend on baseline age, systolic blood pressure, and c-reactive protein level. Using these three variables to stratify the study population leads to a stratified treatment strategy that involves several different combinations of drugs (for patients in different strata).

WNAR Invited 5

The alternative hypothesis: Underrepresented (bio)statisticians share graduate school experiences
Organizer & Chair: Natalie Gasca, California Council on Science and Technology

Underrepresented (bio)statisticians share graduate school experiences
Natalie Gasca (California Council on Science and Technology), Maricela Cruz (Kaiser Permanente Washington Health Research Institute), Aaron Hudson (University of California, Berkeley), Amarise Little (University of Washington), Kyle Conniff (University of California Irvine)

According to NSF data, of all domestic math and statistics PhD doctorates who graduated in 2020 (n=928), only 27% were women, 6% Hispanics/Latinos, 3% Blacks/African Americans, and less than 0.1% American Indians/Alaska Natives. Yet the 2020 census reports 19% Latino, 12% Black, and 0.7% Native Americans. To enhance the representation of diverse and talented individuals in (bio)statistics, it is important to highlight the experiences of those who are navigating these fields, as both a source of inspiration and authenticity. We seek to cultivate an inclusive and supportive community of underrepresented statisticians and allies by sharing the experiences and resilience resources of early career professionals and graduate students who have studied in the WNAR region. We will discuss some of the challenges and benefits that our panelists have experienced as part of historically excluded groups, suggestions for departments to create more supportive environments so that everyone can thrive and succeed in graduate school, and career development tips. Our goal is to share our insights as underrepresented professionals so that students can envision a fulfilling career in (bio)statistics.

WNAR Invited 6

Advancement of network modeling and clustering and applications in omics analysis
Organizer & Chair: Wen Zhou, Colorado State University, Colorado School of Public Health

Bootstrapping Network Data: Conditional versus Marginal Distributions
Keith Levin (University of Wisconsin-Madison), Yichen Qin (University of Cincinnati), Youngser Park (Johns Hopkins University), Carey E. Priebe (Johns Hopkins University), Elizaveta Levina (University of Michigan)*

In network analysis, one frequently must perform inference based upon only one sampled network. This poses a challenge for bootstrap-based approaches, which typically require an iid sample. A class of network models called latent space models overcome these difficulties by generating a network based on unobserved geometric structure, but this raises the question of whether inference in such models should be conducted by conditioning on this latent structure or by marginalizing over it. We develop bootstrap schemes for both cases, i.e., conditional and marginal bootstrap methods for network data. We establish bootstrap validity for both schemes for a broad class of network statistics, including modularity, which has not previously been addressed within the network bootstrap literature. Our experiments include simulated data as well as a thorough exploration of a data set arising from Microsoft Bing search data.

Root and community inference on preferential attachment model of networks
Min Xu (Rutgers University), Harry Crane (Rutgers University)*

We introduce the PAPER (Preferential Attachment Plus Erdős--Rényi) model for random networks, where we let a random network G be the union of a preferential attachment (PA) tree T and additional Erdős--Rényi (ER) random edges. The PA tree component captures the fact that real networks often have an underlying growth process where vertices and edges are added sequentially, while the ER component can be regarded as random noise. Given only a single snapshot of the final network G, we study the problem of constructing confidence sets for the early history, in particular the root node, of the unobserved growth process; the root node can be patient zero in a disease infection or the source of

fake news in a social media network. We propose an inference algorithm based on Gibbs sampling that scales to networks with millions of nodes and provide analysis showing that the expected size of the confidence set is small so long as the noise level of the ER edges is not too large. We also propose variations of the model in which multiple growth processes occur simultaneously, reflecting the growth of many communities, and we use these models to provide a new approach to community detection.

Network Estimation by Mixing: Adaptivity and More

Tianxi Li (University of Virginia), Can M. Le (University of California, Davis)*

Networks analysis has been commonly used to study the interactions between units of complex systems. One problem of particular interest is learning the network,Äôs underlying connection pattern given a single and noisy instantiation. While many methods have been proposed to address this problem in recent years, they usually assume that the true model belongs to a known class, which is not verifiable in most real-world applications. Consequently, network modeling based on these methods either suffers from model misspecification or relies on additional model selection procedures that are not well understood in theory and can potentially be unstable in practice. To address this difficulty, we propose a mixing strategy that leverages available arbitrary models to improve their individual performances. The proposed method is computationally efficient and almost tuning-free; thus, it can be used as an off-the-shelf method for network modeling.

Individual-centered partial information in social networks

Xin Tong (University of Southern California)*

In statistical network analysis, we often assume either the full network is available or multiple subgraphs can be sampled to estimate various global properties of the network. However, in a real social network, people frequently make decisions based on their local view of the network alone. Here, we consider a partial information framework that characterizes the local network centered at a given individual by path length L and gives rise to a partial adjacency matrix. Under L = 2, we focus on the problem of (global) community detection using the popular stochastic block model (SBM) and its degree-corrected variant (DCSBM). We derive general properties of the eigenvalues and eigenvectors from the signal term of the partial adjacency matrix and propose new spectral-based community detection algorithms that achieve consistency under appropriate conditions. Our analysis also allows us to propose a new centrality measure that assesses the importance of an individual's partial information in determining global community structure.

WNAR Invited 7

Innovative Advances in Functional Data Applications

Organizer: Brian Kwan, University of California, Los Angeles
Chair: Jingjing Zou, University of California, San Diego

A Bayesian Covariance Based Clustering for High Dimensional Tensors

Rene Gutierrez Marquez (Texas A&M), Aaron Scheffler (University of California, San Francisco), Rajarshi Guhaniyogi (Texas A&M)*

Clustering of high dimensional tensors with limited sample size has become prevalent in a variety of application areas. Existing Bayesian model based clustering of tensors yields less accurate clusters when the tensor dimensions are sufficiently large, sample size is low and clusters of tensors mainly reveal differences in their variability. This article develops a novel clustering technique for high dimensional tensors with limited sample size when the clusters show difference in their covariances, rather than in their means. The proposed approach constructs several matrices from a tensor to adequately estimate its variability along different modes and implements a model-based approximate Bayesian clustering algorithm with the matrices thus constructed, in place with the original tensor data. Although some information in the data is discarded, we gain substantial computational efficiency and accuracy in

clustering. Simulation study assesses the proposed approach along with its competitors in terms of estimating the number of clusters, identification of the modal cluster membership along with the probability of mis-classification in clustering.

Leveraging Functional Data Analysis Methods for Practical Health Applications

Brian Kwan (University of California, Los Angeles), Catherine Sugar (University of California, Los Angeles), Damla Senturk (University of California, Los Angeles), Jing Zhang (University of California, San Diego), Loki Natarajan (University of California, San Diego)*

Technological advances have yielded diverse types of data that can be characterized as curves, contributing to the growth of both methods for and applications of functional data. For instance, functional principal components analysis detects major modes of curve variation, which can identify the most critical components of disease progression based on sparsely sampled data. We have applied this framework to characterize and predict outcome trajectories and elucidate clinically meaningful subgroups in the context of diabetic kidney disease which has a highly heterogeneous presentation. Functional methods are equally promising for biomarker development in the context of densely sampled hierarchical data such as electroencephalography, which may consist of continuously recorded power spectral densities or event-related potential waveforms recorded across multiple trials and electrodes. Such data are increasingly common in studies of psychiatric illnesses such as autism. In this talk, we will illustrate the application of novel functional techniques from data such as the Chronic Renal Insufficiency Cohort and the Autism Biomarkers Consortium for Clinical Trials.

Adaptive Functional Principal Component Analysis

Angel Garcia de la Garza (Columbia University), Britton Sauerbrei (Case Western Reserve University), Adam Hantman (University of North Carolina), Jeff Goldsmith (Columbia University)*

Recent advances have allowed high-resolution observations of firing rates for a collection of individual neurons; these observations can provide insights into patterns of brain activation during the execution of tasks. Our data come from an experiment in which mice performed a reaching motion following an auditory cue, and contain measurements on firing rates from neuron activation in the motor cortex before and after the cue. In this setting, steep increases in firing rates after the cue are expected. Our dimension reduction technique adequately models these sharp changes over time and correctly captures these activation patterns. Initial results suggest different patterns of activation, representing the involvement of different motor cortex functions at different times in the reaching motion.

Minimax Powerful Functional Analysis of Covariance Tests for Longitudinal Genome-Wide Association Studies

Yehua Li (University of California, Riverside)*

We model the Alzheimer's Disease (AD) related phenotype response variables observed on irregular time points in longitudinal Genome-Wide Association Studies (GWAS) as sparse functional data and propose nonparametric test procedures to detect functional genotype effects, while controlling the confounding effects of environmental covariates. Existing nonparametric tests do not take into account within-subject correlations, suffer from low statistical power, and fail to reach the genome-wide significance level. We propose a new class of functional analysis of covariance (fANCOVA) tests based on a seemingly unrelated (SU) kernel smoother, which can incorporate the correlations. We show that the proposed SU-fANCOVA test combined with a uniformly consistent nonparametric covariance function estimator enjoys the Wilks phenomenon and is minimax most powerful. In an application to the Alzheimer's Disease Neuroimaging Initiative data, the proposed test leads to discovery of new genes that may be related to AD.

WNAR Invited 8

Recent Development in Mobile/Wearable Device Data Analysis

Organizer: Jingjing Zou, University of California, San Diego

Chair: Loki Natarajan, University of California, San Diego

Functional data analysis Inference in crossover designs with application to cat accelerometer data
Salil Koner (Duke University), Ana-Maria Staicu (North Carolina State University), Arnab Maity (North Carolina State University)*

Wearable devices for continuous monitoring of electronic health have increased attention due to their richness in information. Oftentimes inference is drawn from summaries that quantify some feature of the data, leading to a loss of information that could be useful when one utilizes the functional nature of the response. This work is motivated by the interest to understand the efficacy of the meloxicam treatment for Osteoarthritis in household cats, by employing a crossover experiment and using the cats' minute-by-minute activity levels as a proxy objective measure of the cat's condition. We develop a testing procedure to test for significance of the treatment effect in a crossover design, in the presence of carryover effect. We propose an orthogonal projection-based test pseudo generalized F test and study its null asymptotic distribution under mild conditions. In numerical studies, the proposed test maintains the size, is powerful to detect the significance of functional treatment effect, and very efficient compared to bootstrap-based alternatives.

Cumulative head impact exposure and the concussion risk: functional data analytic approach
Jaroslaw Harezlak (Indiana University), Brian Stemper (Medical College of Wisconsin), Alok Shah (Medical College of Wisconsin)*

Sport-related concussion (SRC) is a major public health problem resulting in over 200,000 annual trips to the Emergency Departments in the United States. From a biomechanical standpoint, the concussion mechanism involves head impact resulting in high magnitude head rotational accelerations. Mounting evidence from human studies has demonstrated that repetitive head impact exposure (HIE) contributes to decreased SRC tolerance in contact sport athletes. However, studies focusing on quantifying the relationship between the HIE and incident concussion have suffered from simplistic statistical methods utilized. In our research, we use head impact telemetry (HIT) system accelerometry time series data collected on American Football players in the NCAA-DoD Concussion Assessment, Research and Education Consortium (CARE) study to determine the association of HIE with the SRC occurrence and post-SRC recovery based on the characteristics of head impacts. We utilize a scalar-on-function regression approach to determine the most important head impact features as well as their time-varying influence on the SRC.

A Riemann Manifold Model Framework for Longitudinal Changes in Physical Activity Patterns
Jingjing Zou (University of California, San Diego), Tuo Lin (University of California, San Diego), Chongzhi Di (Fred Hutchinson Cancer Research Center), Loki Natarajan (University of California, San Diego)*

Physical activity (PA) is associated with many health outcomes. The wide usage of wearable accelerometer-based activity trackers has provided a unique opportunity for in-depth research on PA and its relations with health outcomes and interventions. Past analysis relies heavily on aggregating minute-level PA records into day-level summary statistics, in which important information of diurnal PA patterns is lost. We propose a novel functional data analysis approach based on theory of Riemann manifolds for modeling longitudinal changes in PA temporal patterns. We model smoothed minute-level PA of a day as one-dimensional Riemann manifolds and longitudinal changes in PA in different visits as deformations between manifolds. Functional principal components (PC) analysis is adopted to model deformation momenta and PC scores are used as a proxy in modeling the relation between changes in PA and health outcomes and/or interventions. We conduct comprehensive analyses on data from two clinical trials: Reach for Health and Metabolism, Exercise and Nutrition at UCSD.

Longitudinal functional principal component analysis on multi-level physical activity data

Wenyi Lin (University of California San Diego), Jingjing Zou (University of California San Diego), Chongzhi Di (Fred Hutchinson Cancer Research Center), Cheryl Rock (University of California San Diego), Loki Natarajan (University of California San Diego)*

Engaging in physical activity (PA) is key to leading a healthy lifestyle and preventing disease. Sensor devices, e.g., accelerometers, are used for accurately measuring PA. These devices provide minute-level output, which yields a rich framework for analysis, but also poses difficulty with multilevel output. For instance, PA may be measured continuously over 24-hours on multiple days and visits, while a scalar outcome (e.g., insulin) is usually observed only at the individual or visit level. This leads to a discrepancy in numbers of nested levels between the exposure (PA) and outcomes, raising analytic challenges. One approach is to average PA data over days (or visits), so that PA and the outcome have concordant nested levels. However, this ignores PA variation over days (or visits). Also, averaging when there are proportions of missing day- (or visit-) level data differ across individuals, can bias results. In this study, we address the problem of applying regression with imbalanced multi-level functional exposure and scalar outcomes. Multiple simulations are implemented to examine the impact of imbalanced data on prediction and offer guidelines for selecting optimal methods.

WNAR Invited 9

Unmeasured confounding in causal inference: recent advances

Organizer & Chair: Fan Xia, University of Washington

Mediation analysis with unmeasured treatment-induced confounding

Fan Xia (University of Washington), Kwun Chuen Gary (Chan)*

In causal mediation analysis, covariates affected by the treatment or exposure can be a source of confounding between the mediator and the outcome. Like any confounders, the treatment-induced confounders can be mismeasured or unmeasured, leading to invalid causal inference. Even when correctly measured, treatment-induced confounders are especially challenging because it mediates part of the exposure effect while confounding the exposure effect through the mediator. As a result, the identification and estimation of natural direct and indirect effects of the exposure with the presence of treatment-induced confounders deviate from those of the well-studied average treatment effect. In this paper, we use variables associated with the treatment-induced confounders as their proxies to account for confounding in natural direct/indirect effects identification. We develop the semiparametric theory for estimation and propose estimators that are robust to different types of model misspecifications. We use simulation studies to evaluate the performance of the proposed method.

The synthetic instrument method

Dingke Tang (University of Toronto), Dehan Kong (University of Toronto), Linbo Wang (University of Toronto)*

Inferring causal relationships from observational studies is a predominant problem in social, economical and biomedical sciences. Previous causal studies hinge on either observed confounders, or auxiliary variables such as negative controls and instrumental variables. In this paper, we introduce a novel framework that leverages the information contained in multiple causes. Under the commonly used structural equation model and sparsity conditions, we achieve identification of causal parameters. Furthermore, we develop a simple estimation procedure based on a regularization problem with a specifically designed penalty. We illustrate our framework using breast cancer gene expression data from The Cancer Genome Atlas (TCGA).

A Focusing Framework for Testing Bi-Directional Causal Effects with GWAS Summary Data

Sai Li (Institute of Statistics and Big Data, Renmin University of China), Ting Ye (Department of Biostatistics, University of Washington)*

Mendelian randomization (MR) is a powerful method that uses genetic variants as instrumental variables (IVs) to infer the causal effect of a modifiable exposure on an outcome. Although recent years

have seen many extensions of basic MR methods to be robust to certain violations of assumptions, few methods were proposed to infer bi-directional causal relationships, especially for phenotypes with limited biological understandings. The presence of horizontal pleiotropy adds another layer of complexity. In this article, we show that assumptions for common MR methods are often impossible or too stringent in the existence of bi-directional relationships. We then propose a new focusing framework for testing bi-directional causal effects between two traits with possibly pleiotropic genetic variants. Our proposal can be coupled with many state-of-art MR methods. We provide theoretical guarantees on the Type I error and power of the proposed methods. We demonstrate the robustness of the proposed methods using several simulated and real datasets.

Accounting for hidden bias in test-negative design studies of vaccine effectiveness leveraging double negative controls

Xu Shi (University of Michigan)*

The test-negative design (TND) has become a standard approach to evaluate vaccine effectiveness. Despite TND's potential to reduce unobserved differences in healthcare-seeking behavior (HSB) between vaccinated and unvaccinated subjects, it remains subject to various potential biases. First, residual confounding bias may remain due to unobserved HSB, occupation as a healthcare worker, or previous infection history. Second, because selection into the TND sample is a common consequence of infection and HSB, collider stratification bias may exist when conditioning the analysis on testing, which further induces confounding by latent HSB. Third, generalizability of the results to the general population is not guaranteed. In this talk, we present a novel approach to identify and estimate vaccine effectiveness in the general population by carefully leveraging a pair of negative control exposure and outcome variables to account for potential hidden bias in TND studies. We illustrate our proposed method with extensive simulation and an application to COVID-19 vaccine effectiveness using data from the University of Michigan Health System.

WNAR Invited 10

Functional and shape data analysis in imaging studies

Organizer & Chair: Eardi Lila, Department of Biostatistics, University of Washington

A functional data approach to identifying Alzheimer's disease from manifold imaging data

Eardi Lila (University of Washington)*

We introduce a novel framework for the classification of functional data supported on non-linear, and possibly random, manifold domains. The motivating application is the identification of subjects with Alzheimer's disease from their cortical surface geometry and associated cortical thickness map. The proposed model is based upon a reformulation of the classification problem into a regularized multivariate functional linear regression model. This allows us to adopt a direct approach to the estimation of the most discriminant direction while controlling for its complexity with appropriate differential regularization. We apply the proposed method to a pooled dataset from the Alzheimer's Disease Neuroimaging Initiative and the Parkinson's Progression Markers Initiative, and are able to estimate discriminant directions that capture both cortical geometric and thickness predictive features of Alzheimer's Disease, which are consistent with the existing neuroscience literature.

Gaussian Copula Function-on-Scalar Regression in Reproducing Kernel Hilbert Space

Linglong Kong (University of Alberta)*

To relax the linear assumption in function-on-scalar regression, we borrow the strength of copula and propose a novel Gaussian copula function-on-scalar regression. Our model is more flexible to characterize the dynamic relationship between functional response and scalar predictors. Estimation and prediction are fully investigated. We develop a closed form for the estimator of coefficient functions in a reproducing kernel Hilbert space without the knowledge of marginal transformations. Valid,

distirbution-free, finite-sample prediction bands are constructed via conformal prediction. Theoretically, we establish the optimal convergence rate on the estimation of coefficient functions and show that our proposed estimator is rate-optimal under fixed and random designs. The finite-sample performance is investigated through simulations and illustrated in real data analysis.

Statistical Frameworks for Variable Selection with 3D Shapes and High-Resolution Imaging

Lorin Crawford (Microsoft Research)*

The recent curation of large-scale databases with 3D surface scans of shapes has motivated the development of tools that better detect global patterns in morphological variation. Studies which focus on identifying differences between shapes have been limited to simple pairwise comparisons and rely on pre-specified landmarks (that are often known). In this talk, we present SINATRA: a statistical pipeline for analyzing collections of shapes without requiring any correspondences. Our method takes in two classes of shapes and highlights the physical features that best describe the variation between them. We develop a rigorous simulation framework to assess our approach, which themselves are a novel contribution to 3D image and shape analyses. Lastly, as case studies with real data, we use SINATRA to (1) analyze mandibular molars from four different suborders of primates and (2) facilitate the visual identification of structural signatures differentiating between the trajectories of two protein ensembles resulting from molecular dynamics simulations.

Tangent Functional Canonical Correlation Analysis for Densities and Shapes, with Applications to Multimodal Imaging Data

Sebastian Kurtek (The Ohio State University), Min Ho Cho (Inha University), Karthik Bharath (University of Nottingham)*

It is quite common for functional data arising from imaging data to assume values in infinite-dimensional manifolds. Uncovering associations between two or more such nonlinear functional data extracted from the same object across medical imaging modalities can assist development of personalized treatment strategies. We propose a method for canonical correlation analysis between paired probability densities or shapes of closed planar curves, routinely used in biomedical studies, which combines a convenient linearization and dimension reduction of the data using tangent space coordinates. Leveraging the fact that the corresponding manifolds are submanifolds of unit Hilbert spheres, we describe how finite-dimensional representations of the functional data objects can be easily computed, which then facilitates use of standard multivariate canonical correlation analysis methods. We further construct and visualize canonical variate directions directly on the space of densities or shapes. Utility of the method is demonstrated through numerical simulations and performance on a magnetic resonance imaging dataset of glioblastoma multiforme brain tumors.

WNAR Invited 11

Modern statistical learning methods for spatial and imaging data

Organizer & Chair: Lily Wang, George Mason University

Simulation Experiments and Uncertainty Quantification in Remote Sensing

Emily L. Kang (University of Cincinnati)*

Remote sensing data sets produced by NASA and other space agencies are the result of complex algorithms that infer geophysical state from observed radiances using retrieval algorithms. Simulation experiments are important tools to design new observing systems, to evaluate new data assimilation algorithms, to calibrate parameters, and to study uncertainty propagation. This talk will discuss opportunities and challenges involved in such experiments and to advance statistical methodology. Examples will be given of modeling multivariate spatial processes and constructing a statistical emulator of the physical forward model in remote sensing retrieval algorithms used several in NASA missions.

A Scalable Partitioned Approach to Model Massive Nonstationary Non-Gaussian Spatial Datasets

Ben Seiyon Lee (George Mason University), Jae Woo Park (Yonsei University)*

Nonstationary non-Gaussian spatial data are common in many disciplines, including climate science, ecology, epidemiology, and social sciences. Modeling such datasets as stationary spatial processes can be unrealistic since they are collected over large heterogeneous domains (i.e., spatial behavior differs across subregions). Although several approaches have been developed for nonstationary spatial models, these have focused primarily on Gaussian responses. In addition, fitting nonstationary models for large non-Gaussian datasets is computationally prohibitive. We propose a scalable algorithm for modeling such data by leveraging parallel computing in modern high-performance computing systems. We partition the spatial domain into disjoint subregions and fit locally nonstationary models using a carefully curated set of spatial basis functions. Then, we combine the local processes using a novel neighbor-based weighting scheme. Our approach scales well to massive datasets (e.g., 2.7 million samples) and can be implemented in nimble, a popular software environment for Bayesian hierarchical modeling.

Big Imaging Data Learning: A Parallel Solution

Shan Yu (University of Virginia), Guannan Wang (College of William and Marry), Lily Wang (George Mason University), Lei Gao (George Mason University)*

Explosive growth in spatial and spatiotemporal data emphasizes the need for developing new and computationally efficient methods and credible theoretical support tailored for analyzing such large-scale data. Parallel statistical computing has proved to be a handy tool when dealing with big data. However, it is hard to execute the conventional spatial regressions in parallel. In this work, we develop a novel parallel smoothing technique for generalized partially linear spatially varying coefficient models, which can be used under different hardware parallelism levels. Moreover, conflated with concurrent computing, the proposed method can be easily extended to the distributed system. Regarding the theoretical support of estimators from the proposed parallel algorithm, we first establish the asymptotical normality of linear estimators. Secondly, we show that the spline estimators reach the same convergence rate as the global spline estimators. The proposed method is evaluated through extensive simulation studies and an analysis of the US loan application data.

Statistical Inferences on Neuroimaging Data via Deep Neural Networks

Jian Kang (University of Michigan)*

Regression models with varying coefficients have been widely used to study the associations between the massive imaging data and variables of interests. This problem is challenging, due to the ultrahigh dimensionality, the high and heterogeneous level of noise, and the limited sample size of the imaging data. In this talk, I will present a series of varying coefficient models for imaging data analysis where the varying coefficients are constructed through deep neural networks (DNN). Compared with the existing solutions, our methods are more flexible in capturing the complex patterns among the brain activity, of which the noise level and the spatial dependence appear to be heterogeneous across different brain regions. I will discuss the parameter estimation and inference procedures along with the theoretical properties of our proposed methods. I will show that the new methods outperform the existing ones through both simulations and different neuroimaging data examples.

WNAR Invited 12

Bayesian Clinical Trial Designs and Analyses

Organizer: Masataka Taguri, Yokohama City University

Chair: Kentaro Takeda, Astellas

Model calibration approaches for model uncertainty of dose-efficacy relationship in phase I trials for monotherapy of molecularly targeted agents

Hiroyuki Sato (Tokyo Medical and Dental University), Masanao Sasaki (Tokyo Medical and Dental University), Akihiro Hirakawa (Tokyo Medical and Dental University)*

Some targeted agents (TAs) are considered safe and have maximum efficacy at dose level lower than the maximum tolerated dose. The optimal dose of such agents can be determined by evaluating their efficacy and toxicity outcomes, but the dose-efficacy relationship may exhibit non-monotonic patterns, such as initially increasing with the dose level and then plateauing or even decreasing at higher dose levels. To elucidate this relationship, we developed a model-based dose-finding method using the change-point logistic model for a single TA, termed the change-point (CP) method. The recent development of diverse types of targeted agents has prompted the establishment of adaptive dose-finding methods that can account for various dose-efficacy relationships during trials. In this study, we expand the CP method into model calibration methods to account for model uncertainty in dose-efficacy relationship analyses through model selections based on posterior model probability and Bayesian model averaging in phase I trials for a single TA. We also compare the performance of the proposed and rival model-based dose-finding methods through simulation studies.

Constrained hierarchical Bayesian model for latent subgroups in basket trials with two classifiers

Kentaro Takeda (Astellas), Shufang Liu (Astellas), Alan Rong (Astellas)*

The basket trial is a novel trial design that enables the simultaneous assessment of one drug in multiple cancer types. In addition to the cancer types, many recent basket trials contain other classifiers like biomarkers that potentially affect the clinical outcomes. In other words, the treatment effects in those baskets are often categorized by both the cancer types and the levels of other classifiers. In this article, we propose a constrained hierarchical Bayesian model for latent subgroups (CHBM-LS) to deal with potential heterogeneity of treatment effects due to both the cancer type (first classifier) and another classifier (second classifier) in basket trials. Different baskets defined by cancer types and levels of the second classifier are aggregated into latent subgroups. Within each latent subgroup, the treatment effects are approximately exchangeable. The CHBM-LS approach evaluates the treatment effect for each basket while allowing adaptive information borrowing across the baskets. The simulation study shows that the CHBM-LS approach outperforms other approaches with higher statistical power and better-controlled type I error rates under various scenarios.

Dynamic borrowing from multiple historical control data by shrinkage priors

Tomohiro Ohigashi (University of Tsukuba), Kazushi Maruo (University of Tsukuba), Takashi Sozu (Tokyo University of Science), Ryo Sawamoto (The University of Tokyo), Masahiko Gosho (University of Tsukuba)*

Meta-analytic approaches and power priors are often used to incorporate historical controls into the analysis of a current randomized controlled trial. Ohigashi et al. (2022+) proposed a method for incorporating multiple historical controls based on a horseshoe prior, which is a type of shrinkage prior. The method assumes that multiple historical controls could follow a distribution that is potentially biased from the current control, instead of evaluating the overall mean of current and historical controls, as in the meta-analytic approach. In this study, we propose new methods for incorporating multiple historical controls based on shrinkage priors using Dirichlet,ÄìLaplace and spike-and-slab. When the current and historical controls follow the same distribution, the statistical power and effective historical sample size of the methods with shrinkage priors assumed potentially biases are higher than those of other methods. When a few heterogeneous historical controls exist, the average biases in the parameter of interest of the current control using the proposed method with spike-and-slab prior are smaller than those using other methods.

U-BOIN: a utility-based two-stage phase I-II design to identify maximum tolerated dose and optimal biological dose

Yanhong Zhou (The University of Texas MD Anderson UTHealth Graduate School of Biomedical Sciences), J. Jack Lee (The University of Texas MD Anderson Cancer Center), Ying Yuan (The University of Texas MD Anderson Cancer Center)*

We develop a utility-based Bayesian optimal interval (U-BOIN) phase I/II design to find the OBD. We jointly model toxicity and efficacy using a multinomial-Dirichlet model, and employ a utility function to measure dose risk-benefit trade-off. The U-BOIN design consists of two seamlessly connected stages. In stage I, the Bayesian optimal interval (BOIN) design is used to quickly explore the dose space and collect preliminary toxicity and efficacy data. In stage II, in light of accumulating efficacy and toxicity from both stages I and II, we continuously update the posterior estimate of the utility for each dose after each cohort, and use this information to direct the dose assignment and selection. Compared to existing phase I/II designs, one prominent advantage of the U-BOIN design is its simplicity for implementation (with predetermined decision tables). Our simulation study shows that, despite its simplicity, the U-BOIN design is robust and has high accuracy to identify the OBD. The design has been used in drug development recently. A user-friendly software is also freely available to implement this design.

WNAR Invited 13

Novel statistical methods for analyzing neuroimaging data

Organizer: Cai Li, St. Jude Children's Research Hospital
Chair: Yimei Li, St. Jude Children's Research Hospital

Biobank-scale Multi-organ Imaging Genetics: Clinical and Statistical Advances

Bingxin Zhao (Purdue University)*

The UK Biobank's imaging study has scanned the brains and bodies of more than 40,000 participants using magnetic resonance imaging (MRI), making it the world's largest multi-modal imaging database. Meanwhile, several independent studies have also produced publicly available imaging genetic datasets. We integrate, process, and use MRI data from over 50,000 subjects to examine the genetic architecture of the human brain and body. In this talk, I will present new findings on the genetic influences of human brain white matter and its genetic links to a wide range of clinical outcomes, including glioma and stroke. Further, I will discuss the statistical issues we encountered when analyzing these datasets and the possible directions for future multi-organ imaging research. More information about our studies is available at the Brain Imaging Genetics Knowledge Portal (BIG-KP, https://bigkp.org).

Single-index models with functional connectivity network predictors

Caleb Weaver (North Carolina State University), Luo Xiao (North Carolina State University), Martin Lindquist (Johns Hopkins University)*

Functional connectivity is defined as the undirected association between two or more functional magnetic resonance imaging (fMRI) time series. Increasingly, subject-level functional connectivity data have been used to predict and classify clinical outcomes and subject attributes. We propose a single-index model wherein response variables and sparse functional connectivity network valued predictors are linked by an unspecified smooth function in order to accommodate potentially nonlinear relationships. We exploit the network structure of functional connectivity by imposing meaningful sparsity constraints, which lead not only to the identification of association of interactions between regions with the response but also the assessment of whether or not the functional connectivity associated with a brain region is related to the response variable. We demonstrate the effectiveness of the proposed model in simulation studies and in an application to a resting-state fMRI data set from the Human Connectome Project to model fluid intelligence and sex and to identify predictive links between brain regions.

A High-dimensional Mediation Model for a Neuroimaging Mediator

Xiaoqing Wang (University of Michigan), Yimei Li (St. Jude Children's Research Hospital), Arzu Onar-Thomas (St. Jude Children's Research Hospital), Cheng Cheng (St. Jude Children's Research Hospital), Zhaohua Lu (St. Jude Children's Research Hospital)*

Pediatric cancer treatment, especially for brain tumors, can have profound and complicated late effects. A frontline medulloblastoma clinical trial (SJMB03) has collected data, including treatment, clinical, neuroimaging, and cognitive variables. Advanced methods for modeling and integrating these data are critically needed to understand the mediation pathway from the treatment through brain structure to neurocognitive outcomes. We propose an integrative Bayesian mediation analysis approach to model jointly a treatment exposure, a high-dimensional structural neuroimaging mediator, and a neurocognitive outcome and to uncover the mediation pathway. The high-dimensional imaging-related coefficients are modeled via a binary Ising-Gaussian Markov random field prior (BI-GMRF), addressing the sparsity, spatial dependency, and smoothness and increasing the power to detect brain regions with mediation effects.

Heterogeneity Analysis on Multi-state Brain Functional Connectivity and Adolescent Neurocognition

Yize Zhao (Yale University)*

Brain functional connectivity or connectome provides a great potential to explain the neurobiological underpinning of behavioral profiles. Existing connectome-based predictive models link functional connectivity with a behavioral trait without considerations on the heterogeneity in a brain-to behavior relationship, and the information enhancement by integrating connectomes under different cognitive states. In this work, we propose a unified Bayesian model to characterize the heterogeneous relationship between multi-state functional connectivity and a behavior outcome. We also impose nonparametric Bayesian priors to achieve clustering under this supervised learning paradigm. In light of the modular nature, we model the network predictors through stochastic block structures, and simultaneously select sub-network features to define each subtype. A variational EM algorithm is developed to facilitate a posterior computation. We apply our method to establish the impact of functional connectivity under resting and different tasks on the child's fluid intelligence.

WNAR Invited 14

Functional Data Analysis: New Directions and Innovations

Organizer & Chair: Lily Wang, George Mason University

Statistical Inference for Mean Functions of 3D Functional Objects

Yueying Wang (Columbia University Irving Medical Center), Brandon Klinedinst (Iowa State University), Guannan Wang (College of William and Mary), Auriel Willette (Iowa State University), Lily Wang (George Mason University)*

Recently, 3D medical images, such as fMRI and PET, have been attracting researchers' attention due to their ability to provide detailed characterization of brain activity. These 3D complex images are usually collected within the irregular boundary, whereas most existing statistical methods have been focusing on a regular domain. To address this problem, we model the complex data objects as functional data and propose trivariate spline smoothing on tetrahedralizations for estimating the mean functions of 3D functional objects. Motivated by the need for statistical inference for complex functional objects, we present a novel approach for constructing simultaneous confidence corridors to quantify estimation uncertainty. Extension of the procedure to the two-sample case is discussed together with numerical experiments and a real-data application using Alzheimer's Disease Neuroimaging Initiative database.

Prediction Intervals for Multiple Functional Regression

Ryan Kelly (University of Pittsburgh), Kehui Chen (University of Pittsburgh)*

In this talk, we will discuss the application of conformal prediction techniques to the problem of constructing prediction intervals in a multiple functional regression setting. After a short introduction to

the Signature expansion and its favorable properties, a method utilizing this feature set is developed with great modeling flexibility. With minimal assumptions, the resulting algorithm produces a closed form solution for a prediction set with guaranteed coverage. Conditions necessary for efficiency and contiguity of the prediction set are discussed.

Adaptive Frequency Band Analysis for Functional Time Series

Pramita Bagchi (George Mason University), Scott Bruce (Texas A&M University)*

The frequency-domain properties of nonstationary functional time series often contain valuable information. These properties are characterized through its time-varying power spectrum. Practitioners seeking low-dimensional summary measures of the power spectrum often partition frequencies into bands and create collapsed measures of power within bands. However, standard frequency bands have largely been developed through manual inspection of time series data and may not adequately summarize power spectra. In this article, we propose a framework for adaptive frequency band estimation of nonstationary functional time series that optimally summarizes the time-varying dynamics of the series. We develop a scan statistic and search algorithm to detect changes in the frequency domain. We establish theoretical properties of this framework and develop a computationally-efficient implementation. The validity of our method is also justified through numerous simulation studies and an application to analyzing electroencephalogram data in participants alternating between eyes open and eyes closed conditions.

Q-learning with Functional Imaging Features

Xinyi Li (Clemson University), Michael Kosorok (University of North Carolina at Chapel Hill)*

Precision medicine seeks to discover an optimal personalized treatment plan and thereby provide informed and principled decision support, based on the characteristics of individual patients. With recent advancements in medical imaging, it is crucial to incorporate patient-specific imaging features in the study of individualized treatment regimes. We propose a novel, data-driven method to construct interpretable image features which can be incorporated, along with other features, to guide optimal treatment regimes. The proposed method treats imaging information as a realization of a stochastic process, and employs smoothing techniques in estimation. We show that the proposed estimators are consistent under mild conditions. The proposed method is applied to a dataset provided by the Alzheimer's Disease Neuroimaging Initiative.

WNAR Invited 15

Adaptive Design in Clinical Trials

Organizer: Suhwon Lee, University of Missouri
Chair: Jessica Kohlschmidt, Ohio State University

Designs for Joint Estimation of a Target Dose and the Slope of a Dose-Response Function at the Target

Jose Antonio Moler (Universidad Publica de Navarra), Nancy Flournoy (University of Missouri), Fernando Plo (Universidad de Zaragoza), Seung Won Hyun (Tempus Lab)*

Dose-finding experiments aim to estimate the dose having a targeted expected proportion of positive responses by collecting data in the vicinity of this unknown target dose. Such methods are common in many fields. Without loss of generality, we use the language of toxicity studies in phase I clinical trials and assume a monotonically increasing dose-response function. The importance of estimating the slope at the target dose, in addition to the target itself, was recognized long ago in the statistical literature. This slope is of interest because, if large, a small error in the target dose estimate will result in a dose estimate whose toxicity rate is far from the expected proportion of positive responses; whereas with small slopes at the target dose, a large error in the target dose estimate conveys negligible change on its associated toxicity rate. An assessment of these situations is an important component of the study results summary. But there is a conflict between efficient estimation and local data collection. For

instance, the D-optimal design is unacceptable due to ethical or cost considerations. New designs are proposed in this context.

Likelihood-Based Early Stopping Decisions

Nancy Flournoy (University of Missouri - Columbia), Sergey Tarima (Medical College of Wisconsin)*

The use of early stopping rules grows in popularity. For decades, much theoretical effort has focused on characterizing type I error rates under conditions that ensure the normality of parameter estimates. The likelihood factors into components corresponding to each stopping stage; these are the subdensites of interim test statistics described by Armitage et. al (1969) and commonly used to create stopping boundaries [Tarima and Flournoy (2019 Statistical Papers & 2021 Metrika, https://rdcu.be/cyGfb]. However, we illustrate pictorially that sigma fields are not nested when early stopping options are introduced (as assumed by common asymptotic approximations). Yet working directly on the adapted support permits tractable characterizations of events (which in reality are not normally distributed, even asymptotically) with fewer than usual assumptions. To illustrate the usefulness of the likelihood framework, we show that sequential likelihood ratio tests derived from one-parameter exponential family random variables are UMP for any given alpha-spending function with pre-determined stage-specific sample sizes (no asymptotics or normality assumption required).

Two-stage adaptive enrichment design for randomized clinical trials comparing a treatment with a control

Rosamarie Frieri (Department of Statistical Sciences, University of Bologna, Bologna, Italy), William F. Rosenberger (Department of Statistics, George Mason University, Fairfax, VA), Nancy Flournoy (Department of Statistics, University of Missouri, Columbia, MO), Zhantao Lin (Department of Statistics, Data and Analytics, Eli Lilly and Company, Indianapolis, IN)*

When there is some evidence that the effect of a treatment may differ in certain patients' subgroup, clinical trials enrolling a large heterogeneous population could be highly inefficient. Two stage adaptive enrichment designs use the information accrued as the trial progresses to identify the subpopulation benefiting from the new drug. In the case of a continuous biomarker, adopting a bivariate normal model, we provide a complete framework for both the design and the analysis of an adaptive two-stage enriched trial comparing a new treatment with a control. We show how to characterize the potential enriched population by appropriately choosing a biomarker threshold at the end of stage 1 and we test whether a treatment effect exists in the target population at the end of stage 2. In addition, we investigate sample size selection and we provide guidelines for choosing the number of patients to be enrolled in stage 1 and 2. A potential application of our proposal could be when first stage is a phase II trial, in which the biomarker threshold is identified and stage 2 is a phase III trial in which the enrollment is restricted only to the benefiting subpopulation.

Optimal timing for an interim analysis for a primary endpoint under model assumptions

Seung Won Hyun (Tempus Labs)*

In adaptive clinical trials, an interim analysis is often conducted to terminate a trial either for early success or futility. Early stopping of a trial can save the sponsor valuable resources and potentially bring new therapies to patients faster than anticipated. The timing of an interim analysis is crucial in such settings as it balances the trade-off between collecting the right amount of information and cost savings. In this paper, we propose a novel approach for timing an early success interim look. We demonstrate our approach for both single-arm and placebo controlled trials for testing a primary endpoint using time to event data. Under the assumption that the exposure times follows a piecewise exponential distribution, we study how the variance of the primary parameter of interest changes as a function of the timing of the interim analysis. We present a method to search for the optimal interim timing which can minimize the variance of the parameter of interest in some sense.

Recent Development with Complex Event Time Data

Organizer & Chair: Fei Gao, Fred Hutchinson Cancer Research Center

Efficient Estimation of Semiparametric Transformation Model With Interval-Censored Data in Two-Phase Cohort Studies

Fei Gao (Fred Hutchinson Cancer Research Center), K. C. G. Chan (University of Washington)*

Interval sampling and two-phase sampling have both been advocated for studying rare failure outcomes. With few exceptions focusing on specific designs, they are often studied separately in the statistical literature and require different estimation procedures. We consider efficient estimation of interval-censored data collected in a general two-phase sampling design using a localized nonparametric likelihood. An expectation maximization algorithm is proposed by exploiting multiple layers of data augmentation that handle transformation function, interval censoring, and two-phase sampling structure simultaneously. We study the asymptotic properties of the estimators and conduct inference using profile likelihood. We illustrate the performance of the proposed estimator by simulations and an HIV vaccine trial.

Semiparametric proportional hazards model for left-truncated and interval-censored time-to-event outcome using auxiliary and validated data with application to HCHS/SOL data

Noorie Hyun (Kaiser Permanente Washington Health Research Institute), Pamela A. Shaw (Kaiser Permanente Washington Health Research Institute)*

We can easily observe substantial clinical data collection from large health care community studies or electronic health records (EHR) in health systems in this big data era. Data accuracy can vary according to measurement methods. For example, self-reported medical history can include bias, such as recall bias or response bias. In contrast, biomarkers from a laboratory test are less likely to be biased. We are motivated to study what benefit we can gain by augmenting error-prone self-reported and biomarker-based disease diagnoses in regression for time-to-disease onset. The proposed model addresses left-truncation and interval-censoring in time-to-disease onset outcomes. Also, error in self-reported disease diagnosis is corrected by using sensitivity and specificity parameters in the joint likelihood. Comprehensive simulation studies found appealing finite sample properties of the proposed augmenting model, including the smallest mean square error, compared to other models using only either biomarker or self-reported data. The proposed model is applied to the Hispanic Community Health Study/ Study of Latino data to quantify risk factors associated with diabetes onset

Semiparametric regression analysis of partly interval-censored failure time data with application to an AIDS clinical trial

Qingning Zhou (University of North Carolina at Charlotte), Yanqing Sun (University of North Carolina at Charlotte), Peter B. Gilbert (Fred Hutchinson Cancer Research Center)*

Failure time data subject to various types of censoring commonly arise in epidemiological and biomedical studies. Motivated by an AIDS clinical trial, we consider regression analysis of failure time data that include exact and left-, interval-, and/or right-censored observations, which are often referred to as partly interval-censored failure time data. We study the effects of potentially time-dependent covariates on partly interval-censored failure time via a class of semiparametric transformation models that includes the widely used proportional hazards model and the proportional odds model as special cases. We propose an EM algorithm for the nonparametric maximum likelihood estimation and show that it unifies some existing approaches developed for traditional right-censored data or purely interval-censored data. In particular, the proposed method reduces to the partial likelihood approach in the case of right-censored data under the proportional hazards model. We establish that the resulting estimator is consistent and asymptotically normal. In addition, we investigate the proposed method via simulation studies and apply it to the motivating AIDS clinical trial.

Statistical inference for recurrent event processes under shape heterogeneity

Yifei Sun (Columbia University), Ying Sheng (Chinese Academy of Sciences)*

The Cox-type proportional rate models have been the most popular method for recurrent event data analysis. A key assumption of the proportional rate model is that all subjects share the same baseline rate function and the covariate effects act multiplicatively on the rate function. Although facilitating a straightforward rate-ratio interpretation of covariate effects, this assumption implies that covariates do not modify the shape of the rate functions. We fill in the gap in the existing literature by allowing the shape of the rate function to depend on covariates via a functional single index model. Our model encompasses a variety of existing semiparametric recurrent event models as special cases, while retains the interpretability of covariate effects. To estimate the shape of the rate function, we propose a pseudo likelihood approach to eliminate the impact of size. We then project the event count from the follow-up period to the time interval of interest based on the estimated shape, so that general regression methods can be applied to obtain the size parameters. Simulation studies and an analysis of cancer recurrence data were conducted to illustrate the proposed methods.

WNAR Invited 17

Advances and applications in health precision and prediction of events with correlated data

Organizer & Chair: Elizabeth Juarez-Colunga, University of Colorado Anschutz Medical Campus

Using the Generalized Linear (Mixed) Model to Personalize Interventions

Ann A. Lazar (University of California San Francisco)*

Precision health is a fundamental shift in how doctors treat patients. Rather than treating the average patient, precision health personalizes care by treating patients according to their individual characteristics, thus empowering patients and doctors to recommend interventions based on the individual. Using standard precision health methods, we can compare the relationship between an outcome and covariate for two or more intervention groups by evaluating whether the fitted regression curves differ significantly, also known as an intervention-by-covariate interaction or heterogeneity of treatment effects. When they do, thus HTE is present, we can then determine which subgroups of patients benefit most (or least) from the intervention while controlling the family wise error rate. We can also identify whether a particular subject or future subject with certain characteristics (e.g., older subject) will benefit from the intervention. We show that this methodology is suitable for generalized linear (mixed) models (GLM or GLMM) and how this methodology can be applied to real health data.

Identification of extreme clusters in mixed effects models using optimal weighted predictions

Charles McCulloch (UCSF), John Neuhaus (UCSF), Ross Boylan (UCSF)*

Predicted random effects from mixed effects models fit to clustered data are often used to identify extreme clusters such as poorly performing hospitals. We recently proposed optimal weighted predictors that have much lower prediction error for extreme clusters as compared to the usual best predictors. In this talk we propose prediction intervals for our optimal weighted predictors and compare their performance at identifying extreme clusters. Example data predicting walking speed in older adults from a study of osteoarthritis of the knee motivate this work and illustrate the findings.

Predicting events with longitudinal sparse data

Kristen Miller (University of Colorado Anschutz Medical Campus), Jacob Pellinen (University of Colorado Anschutz Medical Campus), Jacqueline French (NYU Grossman School of Medicine), Elizabeth Juarez-Colunga (University of Colorado Anschutz Medical Campus)*

Epilepsy affects approximately 3.4 million Americans, carrying a lifetime risk of around 3%. Providing an accurate assessment of the time course of a given patient's epilepsy is a significant challenge, due to the lack of large-scale studies and the methodologically complicated nature of the seizure outcome. Seizure

events are often infrequent, and when they do occur, the interpatient variability is high. This combination of sparse data and high variability is difficult to address with traditional statistical methodology. The Human Epilepsy Project (HEP) is a 5-year prospective study that collected longitudinal seizure data and several promising biomarkers on 450 newly treated patients with focal epilepsy, making it the largest study in this population. Using the HEP data, we develop a dynamic prediction model for seizure trajectory that addresses the variability and zero-heavy distribution in seizures. This novel model incorporates a mixture sub-model that identifies subgroups of individuals with similar seizure trajectories. The model leverages all available historical seizure, medication, and demographic data to optimize prediction.

WNAR Invited 18

Network-Based and Functional Data Analysis Approaches for OMICS Data

Organizer & Chair: Debmalya Nandy, Department of Biostatistics & Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus

Dynamic Biomarker Discovery for Childhood Obesity Using Functional Data Analysis

Ana Maria Kenney (UC Berkeley)*

Obesity is highly heritable, yet only a small fraction of its heritability has been attributed to genetic variants. These are traditionally ascertained from genome-wide association studies, which utilize samples with tens or hundreds of thousands of individuals for whom a single summary measurement is collected. An alternative approach is to focus on smaller, more deeply characterized samples. Here functional data analysis techniques are used to capitalize on longitudinal growth information from a cohort of children and construct polygenic risk scores through a weighting approach that incorporates the dynamic and joint effects of SNPs. It is shown that genetic variants identified in young children are informative in older children and adults, consistent with early childhood obesity being predictive of obesity later in life. We complement these results with simulations establishing that deeply characterized samples can be just as, if not more, effective than comparable studies with cross-sectional responses. Overall, we demonstrate that careful characterizations and analyses of longitudinal phenotypes can provide increased statistical power to studies with limited sample sizes.

RCFGL: Rapid Condition adaptive Fused Graphical Lasso

Souvik Seal (CU Anschutz Medical Campus)*

Inferring gene co-expression networks is important for understanding gene regulation and pathway activity. These networks are usually undirected graphs where genes are the nodes and an edge represents a significant co-expression relationship. When gene-expression data from multiple conditions are available, joint estimation of networks harnessing shared information across them can significantly increase the power. Condition adaptive fused graphical lasso (CFGL) is one such method for joint estimation of networks that also accounts for condition specificity i.e., retains condition-specific patterns of co-expression which can provide insights into underlying mechanisms activated in a particular condition. However, the current algorithm is prohibitively slow even for a moderate number of genes and can only be employed for a maximum of three conditions. We propose a faster alternative of CFGL known as rapid condition adaptive fused graphical lasso (RCFGL). Along with being computationally feasible, RCFGL can also jointly analyze more than three conditions. We use RCFGL to estimate the gene co-expression networks of three brain regions from a genetically diverse cohort of rats.

NetSHy: Network Summarization via a Hybrid approach leveraging topological properties

Thao Vu (Colorado School of Public Health), Elizabeth Litkowsky (Colorado School of Public Health), Weixuan Liu (Colorado School of Public Health), Katherine Pratte (National Jewish Health), Yonghua Zhuang (Colorado School of Public Health), Katerina Kechris (Colorado School of Public Health)*

Biological networks provide a system level understanding of the underlying cellular processes. They consist of subsets of nodes, known as modules, which are highly interconnected and performing separate functionality. To investigate the association between the identified module and an outcome of interest, a summarization, which best explains the module's behavior, is needed. Conventional approaches obtain such a representation using only network feature profiles. We propose NetSHy, a hybrid approach to reduce network dimension while utilizing topological properties by applying PCA on the combination of the feature profiles and the Laplacian matrix. Rigorous network simulation scenarios varying sizes and sparsity levels show that NetSHy outperforms the conventional approach in recovering the true correlation with observed phenotype and maintaining a higher amount of explained variation when networks are relatively sparse. The robustness of NetSHy is demonstrated by a more consistent correlation with the observed phenotype as sample size decreases. Lastly, GWAS studies on biological networks using NetSHy summarizations identify more significant SNPs than the conventional counterpart.

Large-scale genomic study reveals robust activation of the immune system following advanced Inner Engineering meditation retreat

Vijayendran Chandran (Department of Pediatrics, College of Medicine, University of Florida, 1600 SW Archer Road, Gainesville, Florida, 32610, USA.), Mei-Ling Bermúdez (Department of Pediatrics, College of Medicine, University of Florida, 1600 SW Archer Road, Gainesville, Florida, 32610, USA.), Mert Koka (Department of Pediatrics, College of Medicine, University of Florida, 1600 SW Archer Road, Gainesville, Florida, 32610, USA.), Brindha Chandran (Department of Pediatrics, College of Medicine, University of Florida, 1600 SW Archer Road, Gainesville, Florida, 32610, USA.), Dhanashri Pawale (Department of Anesthesia, Indiana University School of Medicine, 1130 West Michigan St., Fesler Hall 204, Indianapolis, 46202, USA.), Ramana Vishnubhotla (Department of Anesthesia, Indiana University School of Medicine, 1130 West Michigan St., Fesler Hall 204, Indianapolis, 46202, USA.)*

The positive impact of meditation on human wellbeing is well documented, yet its molecular mechanisms are incompletely understood. We applied a comprehensive systems biology approach starting with whole blood gene expression profiling combined with multi-level bioinformatic analyses to characterize and identify meditation-specific core network after an advanced 8-day Inner Engineering retreat program. We found the response to oxidative stress, detoxification, and cell cycle regulation pathways were downregulated after meditation. Strikingly, 220 genes directly associated with immune response, including 68 genes related to interferon signaling were upregulated, with no significant expression changes in the inflammatory genes. This robust meditation-specific immune response network is significantly dysregulated in multiple sclerosis and severe COVID-19 patients. The work provides a foundation for understanding the effect of meditation and suggests that meditation as a behavioral intervention can voluntarily and non-pharmacologically improve the immune response for treating various conditions associated with excessive inflammation with a dampened immune system profile.

WNAR Invited 19

Advances in Single-Cell Multi-omics Analysis

Organizer: Lingling An, University of Arizona
Chair: Hongmei Jiang, Northwestern University

MIRA: Joint regulatory modeling of multimodal expression and chromatin accessibility in single cells

Clifford Meyer (DFCI and Harvard T.H. Chan School of Public Health), Allen Lynch (Dana-Farber Cancer Institute), Christina Theodoris (DFCI and Boston Children's Hospital), Henry Long (Dana-Farber Cancer Institute), Myles Brown (DFCI and Harvard Medical School), X. Shirley Liu (DFCI and HSPH)*

Single cell chromatin accessibility, measured by ATAC-seq, and single cell gene expression, measured by RNA-seq provide two views on a cell’s state. From chromatin accessibility it is possible to infer

information about the regulation of genes such as the location of putative cis-regulatory elements near a gene and the transcription factors that bind to these regions. Rigorously comparing gene expression and chromatin accessibility in the same single cells could illuminate the logic of how coupling or decoupling of these mechanisms regulates cell fate commitment. MIRA: Probabilistic Multimodal Models for Integrated Regulatory Analysis, is a comprehensive methodology that systematically contrasts transcription and accessibility to infer the regulatory circuitry driving cells along developmental trajectories. MIRA leverages topic modeling of cell states and regulatory potential modeling of individual gene loci. MIRA thereby represents cell states in an efficient and interpretable latent space, infers high fidelity lineage trees, determines key regulators of fate decisions at branch points, and exposes the variable influence of local accessibility on transcription at distinct loci.

Model-free prediction test with application to genomics data

Zhanrui Cai (Carnegie Mellon University), Jing Lei (Carnegie Mellon University), Kathryn Roeder (Carnegie Mellon University)*

Testing the significance of prediction in a regression model is one of the most important topics in statistics. This problem is especially difficult without any parametric assumptions on the data. This paper aims to test the null hypothesis that, given confounding variables Z, X does not significantly contribute to the prediction of Y under the model-free setting, where X and Z are possibly high dimensional. We propose a framework that first fits nonparametric regression models on the Y|X and Y|(X, Z), then compares the prediction power of the two models. The proposed method allows us to leverage the strength of the most powerful regression algorithms developed from the modern machine learning community. The p-value for the test can be easily obtained by permutation. In simulations, we find that the proposed method is more powerful compared to existing methods. The proposed method allows us to draw biologically meaningful conclusions from two genomic data analyses without strong distributional assumptions: (a) testing prediction power of sequencing RNA for the proteins in CITE-seq data, and (b) identification of spatially variable genes in spatially resolved transcriptomics data.

Computational principles and statistical challenges in single-cell data integration

Lingling An (University of Arizona), Zhuo Chen (University of Arizona), Xiang Zhang (University of Arizona)*

The advance of single-cell multimodal assays provides us a powerful tool for investigating multiple dimensions of cellular heterogeneity, enabling new insights into development, tissue homeostasis and disease. A key challenge in the analysis of single-cell multimodal data is to devise appropriate strategies for tying together data across different modalities, such as RNA expression, DNA methylation, chromatin accessibility, protein abundance, gene perturbation, and spatial information. Although existing integration strategies exploit similar mathematical ideas, they typically have distinct goals and rely on different principles and assumptions. Here we present the computational principles, statistical challenges, and future trends in single-cell multi-omics modeling and analyses.

JSNMF enables effective and accurate integrative analysis of single-cell multiomics data

Zhixiang Lin (Department of Statistics, The Chinese University of Hong Kong), Yuanyuan Ma (School of Computer & Information Engineering, Anyang Normal University), Zexuan Sun (Department of Statistics, The Chinese University of Hong Kong), Pengcheng Zeng (Department of Statistics, The Chinese University of Hong Kong), Wenyu Zhang (Department of Statistics, The Chinese University of Hong Kong)*

The single-cell multiomics technologies provide an unprecedented opportunity to study the cellular heterogeneity from different layers of transcriptional regulation. However, the datasets generated from these technologies tend to have high levels of noise, making data analysis challenging. Here, we propose JSNMF, which is a versatile toolkit for the integrative analysis of transcriptomic and epigenomic data profiled from the same cell. The core of JSNMF is an unsupervised method based on non-negative matrix factorization, where it assumes different latent variables for different molecular modalities, and integrates the information of transcriptomic and epigenomic data with consensus graph fusion, which

better tackles the distinct characteristics and levels of noise across different molecular modalities. We applied JSNMF to single-cell multiomics datasets from different tissues and different technologies. The results demonstrate the superior performance of JSNMF in clustering and data visualization of the cells. JSNMF also provides rich biological insight on the markers, cell-type-specific region,Äìgene associations and the functions of the identified cell subpopulation.

WNAR Invited 20

Recent development of statistical methodologies in microbiome data analysis
Organizer: Hongmei Jiang, Northwestern University
Chair: Lingling An, University of Arizona

DeepLINK: Deep learning inference using knockoffs with applications to genomics
Zifan Zhu (University of Southern California), Yingying Fan (University of Southern California), Yinfei Kong (California State University Fullerton), Jinchi Lv (University of Southern California), Fengzhu Sun (University of Southern California)*

Although practically attractive with high prediction and classification power, complicated learning methods often lack interpretability and reproducibility, limiting their scientific usage. A useful remedy is to select truly important variables contributing to the response of interest. We develop a method for deep learning inference using knockoffs, DeepLINK, to achieve the goal of variable selection with controlled error rate in deep learning models. We show that DeepLINK can also have high power in variable selection with a broad class of model designs. DeepLINK is applicable to a broad class of covariate distributions described by the possibly nonlinear latent factor models. The empirical performance of DeepLINK is investigated through extensive simulation studies, where it is shown to achieve FDR control in feature selection with both high selection power and high prediction accuracy. We then apply DeepLINK to three real datasets related to human gut microbiome and murine and human single cell RNA-seq data sets and produce statistical inference results with both reproducibility and biological meanings.

Differential abundance analysis to identify functionally different microbes
Michael Sohn (University of Rochester), Robert Beblavy (University of Rochester), Cynthia Monaco (University of Rochester)*

Identifying differentially abundant (DA) microbes between different populations (e.g., healthy vs. diseased) is one of the main goals in the human microbiome study. There are many computational and statistical methods for this type of analysis. However, these currently available methods do not take functional information into account despite the consensus that multiple microbes perform the same or very similar functions. Thus, some of the identified microbes by these methods might not be relevant to the traits of a population. For instance, if microbes A and B perform the same function, the difference in their abundance between the healthy and the diseased is unlikely related to the phenotypes of a disease (even though they might be used for an indication of the disease). To overcome this problem, we propose a two-stage statistical model that identifies DA microbes conditional on their functional differences, thus selecting microbes functionally relevant to the traits of a population.

Microbial Interactions and Community Stability from Longitudinal Microbiome Study
Linchen He (Novartis Pharmaceuticals Corporation), Huilin Li (New York University)*

Dynamic changes of microbiome communities may play important roles in human health and diseases. The recent rise in longitudinal microbiome studies calls for statistical methods that can model the temporal dynamic patterns and simultaneously quantify the microbial interactions and community stability. Here, we propose a novel autoregressive zero-inflated mixed-effects model (ARZIMM) to capture the sparse microbial interactions and estimate the community stability. ARZIMM employs a zero-inflated Poisson autoregressive model to model the excessive zero abundances and the non-zero

abundances separately, a random effect to investigate the underlining dynamic pattern shared within the group, and a Lasso-type penalty to capture and estimate the sparse microbial interactions. Based on the estimated microbial interaction matrix, we further derive the estimate of community stability, and identify the core dynamic patterns through network inference. Through extensive simulation studies and real data analyses we evaluated ARZIMM in comparison with the other methods.

Cluster Analysis of Longitudinal Profiles for Compositional Count Data to Study the Competition-Colonization Trade-Off in Ecology

Chenyang Duan (Oregon State University), Yuan Jiang (Oregon State University)*

We propose a novel method named COMPARING for cluster analysis of longitudinal profiles for the species abundances in a biological system. In COMPARING, GEE 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 further develop and implement the L-ADMM algorithm to estimate the parameters in the proposed model and establish its convergence property. Theoretically, we establish the asymptotic convergence rate of the estimated curves to the true curves and conclude that the subgroups can be correctly identified with a high probability. Empirically, we use simulation studies to show the advantage of COMPARING over its competitors in terms of the accuracy of recovering the underlying clusters of longitudinal trajectories. In addition, we apply COMPARING to reveal the co-existence of blood-borne parasites in African buffalo and demonstrate how the method successfully detects biologically meaningful subgroups of parasites for the competition-colonization trade-off.

WNAR Invited 21

Emerging Challenges and Opportunities in Statistics for Higher-Education and Organizations

Organizer & Chair: James Molyneux, Oregon State University - Department of Statistics

Assessing the Effect of Scholarships on Student Success at Oregon State University

Claudio Fuentes (Oregon State University), Njesa Totty (Oregon State University), James Molyneux (Oregon State University)*

In this talk we look the relationship between graduation/retention rates (as measures of student success) and amount of financial aid for students of different demographic groups at Oregon State University. Using logistic regression models we are able to characterize this relationship and quantify the effect of financial aid on student success while accounting for group differences determined by demographic variables of interest.

Teaching the "Right" Data Science and Statistics to an Audience

James Molyneux (Oregon State University Department of Statistics)*

As the era of "Big Data" has progressed, the number and types of roles requiring knowledge of data science and statistics has continued to expand. And with this expansion has come the realization that the methods that are taught, and indeed how those methods are taught, depends in large part on the intended audience. Since a "one size fits all" set of data science skills and teaching methods does not exist, considerations need to be made in order to teach the "right" set of data science and statistics skills to the intended audience. Using our experience, and based on a set of examples, accrued from teaching high school math and science teachers topics in data science, we offer a set of lessons learned and other considerations which make the teaching of statistics and data science skills more effective. We discuss prioritizing ideas based on time restrictions and how to adapt data science skills to suit the needs of, potentially, non-data science practitioners.

Enhancing the Teaching of Computational Methods for Undergraduate Students

Njesa Totty (Oregon State University), James Molyneux (Oregon State University), Claudio Fuentes (Oregon State University)*

Statistical computing methods are progressively appearing in the textbooks and curricula of courses that introduce undergraduate students to statistical methods. As such, the continual assessment of how these methods are taught and what students should learn is needed. In this talk we address issues surrounding the teaching of simple bootstrap-based methods for hypothesis testing and uncertainty quantification. We discuss the importance of communicating the assumptions behind these methods in order to ensure proper use and enhance the learning experience of students. Through simulation we observe non-trivial differences in their performance when their assumptions are and are not met. These differences further emphasize the importance of discussing these assumptions with students. Therefore, we also discuss an R package, which can be used to introduce undergraduate students to these bootstrap methods, while emphasizing their assumptions.

Scaling the Teaching of Statistics and Data Science for Varied Learning Modes

Vik Gopal (National University of Singapore)*

"The statistics curriculum today is very different from what it was 20 years ago. It differs in terms of delivery modes, assessment and content. Some of these differences are due to improvements in the technology we have, but many changes are due to the change in expectations of what a practicing statistician or data scientist needs to be able to do. Further changes have been forced upon us by *that* virus. One of the biggest challenges I face today is coming up with good assessments, that will not be undermined by Google, Stackoverflow, or collusion. Another challenge is to cater to large classes - large classes that might be physically present, on Zoom, or a combination of these two. Moreover, some classes are delivered through traditional lectures, while others are delivered as "blended" classes. Yet others are meant for self-paced learning. Of necessity, these different learning modes demand custom tools for delivery. Facing these challenges requires quite some deliberation on what to teach, and how to teach it. In this talk, I would like to share my thoughts and ideas about solving some of these challenges and surviving in this fast-paced teaching environment."

WNAR Invited 22

Recent advances in clinical drug development

Organizer: Ting Ye, University of Washington
Chair: Yanyao Yi, Eli Lilly and Company

Evaluation of machine learning approaches for estimating individualized treatment regimens for time-to-event outcomes in observational studies

Ilya Lipkovich (Eli Lilly), Zbigniew Kadziola (Eli Lilly), Duzhe Wang (Eli Lilly), Douglas Faries (Eli Lilly)*

In this presentation we provide an overview and evaluation of machine learning methods for estimating individualized treatment regimens (ITR) for time-to-event outcomes maximizing restricted mean survival time (RMST) in observational studies with non-randomized treatment assignment. We present extensive simulation studies that closely mimicked real-world data under a set of scenarios representing different degrees of alignment between the observed regimen, which reflects the actual prescribing practice, and the optimal ITR. The simulation results include performance characteristics of the candidate methods in terms of their ability to recover the true optimal ITR and various empirical measures of RMST gain based on the comparison between the estimated ITR and the actual prescribing practice.

Application of AI to Clinical Trials

Qi Tang (Sanofi)*

AI has made significant inroads into drug discovery with several AI-developed drugs already in the clinical testing stage. However, the adoption of AI in clinical trials is relatively slow due to small data set and stringent regulatory policies. In this talk, we will examine current landscape of application of AI in

clinical trials, highlight potential low hanging fruits and demonstrate a case study for responder identification.

Inference on Selected Subgroup: One Subgroup, Two Trials, and Three Evaluations

Xinzhou Guo (Hong Kong University of Science and Technology), Jianjun Zhou (Yunnan University), Xuming He (University of Michigan)*

When a promising subgroup is identified from an unsuccessful trial with a broad target population, we often need to evaluate and possibly confirm the selected subgroup with a follow-up validation trial. A direct evaluation of the subgroup from the subjects in both trials is not recommended because of the risk of data snooping. An evaluation based solely on the validation trial is free of bias, but does not make full use of the data in the earlier trial. We show that it is possible to utilize data from both trials to improve the efficiency of post-selection subgroup evaluation. In particular, we propose a new resampling-based approach to quantify and remove selection bias and then to perform data combination from both trials for valid and efficient inference on selected subgroup. The proposed method is model-free and asymptotically sharp. We demonstrate the merit of the proposed method by revisiting the panitumumab trial and show how much data combination could help improve efficiency of clinical trials when a promising subgroup is identified post hoc from part of the data.

Probability of Study Success (PrSS) assessment based on Bias-adjusted Bayesian Network Meta-Analysis (BaBNMA)

Ying Zhang (Eli Lilly), Karen Liu (Eli Lilly), Lei Shen (Eli Lilly), Jingyi Liu (Eli Lilly)*

Drug development is a complex and costly process, and one of the most critical decision is if a compound should move to late-stage clinical trial for registration purpose once the proof of concept study result is available. One big challenge is the effectiveness of the drug is unknown and the treatment effect is not easy to quantify. Bayesian Network Neta-Analysis (BNMA) is commonly used to synthesize relevant data and derive the distribution of multiple treatment effects, which is essential for Probability of Study Success (PrSS) calculation. Then PrSS could be used to facilitate the investment decision. However, selection bias exists since only compounds with promising results from the small study (Ph2) would move to the next development stage (Ph3). The BNMA with selection-bias adjustment has been used to address this issue. A hierarchical model for distribution of the efficacy for a portfolio of compounds was considered, which can serve as the prior distribution. The BaBNMA approach was illustrated via a hypothetical late stage compound where the triggered Go or NO-GO decision making need to be made.

WNAR Invited 23

Novel borrowing approaches using real world data for clinical trials

Organizer & Chair: Herb Pang, Genentech

Innovative clinical trial designs with hybrid control: challenges and opportunities

Jiawen Zhu (Genentech)*

There has been increasing interest in leveraging existing external control data, especially real world data to augment a clinical trial. In recent years, the quality and availability of real world data have been largely improved. However, information borrowing by directly pooling such external controls with study controls may lead to biased estimates of the treatment effect. A hybrid control approach has been proposed, which allows collection of randomized data, and utilizing external control at the same time. Dynamic borrowing methods under the Bayesian framework have been proposed to better control the false positive error through adaptively down-weighting external control data when the two control sources are different. In the talk, we will review recent method development in the hybrid control space, and discuss opportunities and challenges with examples.

Evaluation of Statistical Methods for Hybrid External Controls in Clinical Trials

Xiang Zhang (CSL Behring)*

Single arm trials supplemented by external control arms have been frequently used for drug development in rare diseases and scenarios where randomization to a control group is unethical/unfeasible. However, the validity of any decision making based on such design depends heavily on the appropriateness and quality of the control arm data. FDA guidance lists multiple bias-generating concerns with the use of external controls arising from data quality and validity issues. Hybrid control designs, where clinical trials with a treatment group and a small underpowered control group supplemented with external control data, have the potential to address some of these bias concerns. In this presentation we will introduce several methods that could help analyze the data from such hybrid control design and present a simulation study to evaluate the operating characteristics of single and hybrid real-world control methods across the bias-generating scenarios mentioned in FDA guidance.

Theory and Application of Integrative Analysis of Randomized Clinical Trials with Real-World Data

Xiaofei Wang (Duke University), Dasom Lee (NC State University), Shu Yang (NC State University)*

In this talk, we exploit the complementing features of randomized clinical trials (RCT) and real world evidence (RWE) to estimate the average treatment effect of the target population. We will review existing methods in conducting integrated analysis of binary, continuous and survival data from RCTs and RWEs. We will then discuss in detail new calibration weighting estimators that are able to calibrate the covariate information between RCTs and RWEs. We will briefly review asymptotic results under mild regularity conditions, and confirm the finite sample performances of the proposed estimators by simulation experiments. In a comparison of existing methods, we illustrate our proposed methods to estimate the effect of adjuvant chemotherapy in early-stage resected non,Äìsmall-cell lung cancer integrating data from a RCT and the National Cancer Database.

WNAR Invited 24

Statistical modeling, forecasting and optimal designing of patient enrollment under various restrictions in multicenter clinical trials

Organizer: Mitchell Schepps, University of California, Los Angeles
Chair: Weng Kee Wong, University of California, Los Angeles

Bayesian Accrual Modeling and Prediction in Multicenter Clinical Trials with Varying Center Activation Times

Junhao Liu (Novartis Pharmaceuticals), Yu Jiang (University of Memphis), Byron Gajewski (University of Kansas Medical Center)*

Investigators who manage multicenter clinical trials need to pay careful attention to patterns of subject accrual, and the prediction of activation time for pending centers is potentially crucial for subject accrual prediction. We propose a Bayesian hierarchical model to predict subject accrual for multicenter clinical trials in which center activation times vary. We define center activation time as the time at which a center can begin enrolling patients in the trial. The difference in activation times between centers is assumed to follow an exponential distribution, and the model of subject accrual integrates prior information for the study with actual enrollment progress. We apply our proposed Bayesian multicenter accrual model to two multicenter clinical studies. In summary, the Bayesian multicenter accrual model provides a prediction of subject accrual while accounting for both center- and individual patient-level variation.

Multiple-objective metaheuristic optimization algorithms for complex clinical trial enrollment designs.

Mitchell Schepps (UCLA Department of Biostatistics, Los Angeles, CA), Weng Kee Wong, Los Angeles, CA (UCLA Department of Biostatistics), Matt Austin (Center for Design and Analysis, Amgen Inc. Thousand Oaks, CA), Vladimir Anisimov (Center for Design and Analysis, Amgen Ltd., London, UK)*

Multi-center Clinical trials enroll a large number of participants from multiple centers and frequently they may not meet the target number on time due to a myriad of administrative, cost and unanticipated issues. Some attempts have been used to formalize enrollment strategies using a statistical model and optimization techniques. The Poisson-Gamma mixture regression model is a popular statistical tool to predict and track enrollment rate, and make inference. For multi-center trials conducted globally, the optimization problem is much more complex, and may include multiple objectives and nonlinear constraints that arise from physical, political or budgetary considerations, and additional practical limits imposed by the various centers in different countries. Standard optimization techniques and mathematical programming methods do not usually work well for solving such large complex optimization problems. Nature-inspired metaheuristic algorithms have emerged as a powerful and dominating optimization tools in computer science and engineering to solve complex and high-dimensional optimization problems.

Statistical modeling of restricted enrollment and optimal cost-efficient design in multicenter clinical trials
Vladimir Anisimov (Data Science, Center for Design & Analysis, Amgen Ltd., London, UK), Matthew Austin (Data Science, Center for Design & Analysis, Amgen Inc., Thousand Oaks, CA, USA)*

Design and forecasting patient enrollment are among the greatest challenges that clinical research enterprises face today. The talk is describing the innovative developments in the statistical methodology for modeling and forecasting patient enrollment. The underlying technique uses a Poisson-gamma model developed by Anisimov & Fedorov. New analytic techniques for modeling country enrollment process by a Poisson-gamma process with aggregated parameters and for modeling enrollment under some restrictions (caps on enrollment in countries) are developed. Some discussion on using historic data to improve the prediction of enrollment in the new trials is provided. These results are used for solving the problem of optimal cost-efficient enrollment design: find an optimal allocation of sites/countries that minimizes the global trial cost given that the probability to reach an enrollment target in time is no less than some prescribed probability. Different techniques to find an optimal solution for high dimensional optimization problem for the cases of unrestricted and restricted enrollment and for a small and large number of countries are discussed.

Optimal design of experiments with the observation censoring driven by random enrollment of subjects
Xiaoqiang Xue (Syneos Health), Valerii Fedorov*

In the typical clinical trials the subject arrival times can be modeled as the outcomes of a point process. Subsequently the follow-up times can be viewed as random variables and they become known only after trial completion. Thus the information that is gained depends on the sample size (number of subjects), enrollment process and random follow up times (specific for each subject). We present results for proportional hazard model that follows from optimal design theory and illuminate them with numerical examples.

WNAR Invited 25

What's new in spatial statistics?
Organizer & Chair: Weining Shen, University of California, Irvine

BAMDT: Bayesian Additive Partial Multivariate Decision Trees for Nonparametric Regression
Zhao Tang Luo (Texas A&M University), Huiyan Sang (Texas A&M University), Bani Mallick (Texas A&M University)*

Bayesian additive regression trees (BART) have gained great popularity as a flexible nonparametric function estimation and modeling tool. In this work, we develop a new class of Bayesian additive multivariate decision tree models that combine univariate split rules for handling possibly high dimensional features without known multivariate structures and novel multivariate split rules for features with multivariate structures in each weak learner. The proposed multivariate split rules are

built upon a stochastic bipartition model to achieve highly flexible nonlinear decision boundaries on manifold feature spaces while enabling efficient dimension reduction computations. We demonstrate the superior performance of the proposed method over BART and Gaussian process regression models using simulation data and a Sacramento housing price data set.

Mapping the Prevalence of Demographic and Health Indicators Using Household Surveys

Geir-Arne Fuglstad (Norwegian University of Science and Technology), Zehang Li (University of California, Santa Cruz), Jon Wakefield (University of Washington)*

The emerging need for subnational estimation of demographic and health indicators in low- and middle-income countries (LMICs) is driving a move from design-based methods to spatial and spatio-temporal approaches. The latter are model-based and over- come data sparsity by borrowing strength across space, time and covariates and can, in principle, be leveraged to create yearly fine-scale pixel level maps based on household surveys. However, typical implementations of the model-based approaches do not fully acknowledge the complex survey design, and do not enjoy the theoretical consistency of design-based approaches. We describe how spatial and spatio-temporal methods are currently used for small area estimation in the context of LMICs, highlight the key challenges that need to be overcome, and discuss a new approach, which is methodologically closer in spirit to small area estimation. The main discussion points are demonstrated through case studies using Demographic and Health Surveys (DHS) in Malawi.

Semi-parametric estimation of spatially indexed probability densities with application to regional soil erosion assessment

Zhengyuan Zhu (Iowa State University)*

Local erosion distribution within small watershed is important for erosion management, yet difficult to obtain without costly field work at targeted location. Adopting recent development in analyzing probability density objects utilizing the Wasserstein space, we transform local erosion distributions to squared integrable trajectories, basis expansion of which further isolate spatial structure into multivariate random fields that can be handled by existing geostatistics tools with some generalization, enabling flexible modeling and prediction of these spatially distributed distribution objects. Simulation suggests that the proposed method performs comparably to parametric methods based on correct models, and outperforms conventional approaches when there is model misspecification. Over Shaanxi province of China, based on existing limited survey data, detailed local erosion profile is obtained conditioning on local land use type and covariates derived from digital elevation model.

WNAR Invited 26

New Developments of Bayesian modeling and applications

Organizer & Chair: Hsin-Hsiung Huang, University of Central Florida

Sparse Bayesian Matrix-variate Regression with High-dimensional Binary Response Data

Hsin-Hsiung Huang (University of Central Florida), Shao-Hsuan Wang (National Central University), Qing He (University of Central Florida), Charles Harrison (University of Central Florida), Teng Zhang (University of Central Florida), Jie Yang (University of Illinois Chicago)*

We propose a Bayesian generalized linear models for matrix-valued covariates data with shrinkage priors for estimation and variable selection in high-dimensional settings where the dimensions of the covariates increase as the sample size increases. This study is motivated by extending Bayesian approaches in a classical multivariate linear model. The proposed estimation can be applied to classifying matrix data such as images. We show that the proposed model achieves strong posterior consistency when the dimension grows at a sub-exponential rate with the sample size. Furthermore, we quantify the posterior contraction rate at which the posterior shrinks around the true regression

coefficients. Simulation studies and an application to Electroencephalography and Leucorrhea data show the superior performance of the proposed method over the existing approaches.

Analyzing Zero-inflated Data

Jie Yang (University of Illinois at Chicago)*

Count data with a large portion of zeros arise naturally in many scientific disciplines, including security, insurance, health care, microbiome, and more. When conducting one-sample Kolmogorov-Smirnov (KS) test for count data, the estimated p-value is biased due to plugging in sample estimates of unknown parameters. As a consequence, the result of a KS test could be too conservative. In the newly developed R package iZID for zero-inflated count data, we use bootstrapped Monte Carlo estimates to overcome the bias issue in estimating p-values, as well as bootstrapped likelihood ratio tests for zero-inflated model selection. Our package also provides miscellaneous functions to simulate zero-inflated count data and calculate maximum likelihood estimates of unknown parameters. Compared with other R packages available so far, our package covers more types of zero-inflated distributions and provides adjusted p-value estimates after incorporating the influence of unknown model parameters.

Ultra-high Dimensional Bayesian Matrix-variate logistic regression Variable Selection

Shao-Hsuan Wang (National Central University, Taiwan)*

We propose a Bayesian generalized linear model for matrix-valued covariates data with shrinkage priors for estimation and variable selection in high-dimensional settings where the dimensions of the covariates increase as the sample size increases. This study is motivated by extending Bayesian approaches in a classical multivariate linear model. The proposed estimation can be applied to classifying matrix data such as images. We show that the proposed model achieves strong posterior consistency when the dimension grows at a subexponential rate with the sample size. Furthermore, we quantify the posterior contraction rate at which the posterior shrinks around the true regression coefficients. Simulation studies and an application to Electroencephalography and Leucorrhea data show the superior performance of the proposed method over the existing approaches.

Robust Regularized Low-Rank Matrix Models

Teng Zhang (University of Central Florida)*

While matrix variate regression models have been studied in many existing works, classical statistical and computational methods for the analysis of the regression coefficient estimation are highly affected by high dimensional and noisy matrix-valued predictors. To address these issues, this paper proposes a framework of matrix variate regression models based on a rank constraint, vector regularization (e.g., sparsity), and a general loss function with three special cases considered: ordinary matrix regression, robust matrix regression, and matrix logistic regression. We also propose an alternating projected gradient descent algorithm. Our theoretical analysis can be applied to general optimization problems on manifolds with bounded curvature and can be considered an important technical contribution to this work. We validate the proposed method through simulation studies and real image data examples.

WNAR Invited 27

The New England Journal of Statistics in Data Science (NEJSDS) Invited Papers on Innovative Clinical Trial Designs

Organizer & Chair: Colin Wu, National Heart, Lung and Blood Institute, National Institutes of Health

Flexible Conditional Borrowing Approaches for Leveraging Historical Data in the Bayesian Design of Superiority Trials

Wenlin Yuan (University of Connecticut), Ming-Hui Chen (University of Connecticut), John Zhong (REGENXBIO Inc.)*

In this paper, we consider the Bayesian design of a randomized, double-blind, placebo-controlled superiority clinical trial. To leverage multiple historical data sets to augment the placebo-controlled arm, we develop three conditional borrowing approaches built upon the borrowing-by-parts prior, the hierarchical prior, and the robust mixture prior. The operating characteristics of the conditional borrowing approaches are examined. Extensive simulation studies are carried out to empirically demonstrate the superiority of the conditional borrowing approaches over the unconditional borrowing or no-borrowing approaches in terms of controlling type I error, maintaining good power, having a large ``sweet-spot'' region, minimizing bias, and reducing the mean squared error of the posterior estimate of the mean parameter of the placebo-controlled arm. Computational algorithms are also developed for calculating the Bayesian type I error and power as well as the corresponding simulation errors.

A Novel Design for Establishing Mixed Chimerism and Dose-finding in Transplantation and Cancer Trials

Jiying Zou (Stanford University), John Tamaresis (Stanford University), Robert Lowsky (Stanford University), Ying Lu (Stanford University)*

The main limitations to successful and safe organ transplantation are immune-mediated graft rejection and medical comorbidities induced by the combination of immune suppression (IS) medications taken by the recipient to deter rejection. a single very low dose of total body irradiation (svldTBI) that conditioning the recipient may result in mixed chimerism to support IS drug tapering, minimization, and cessation while maintaining normal graft function. We propose a modified probability interval (mPI) design for finding the biologically optimal dose. We also propose a novel 3+3 rule-based heuristic dose-selection algorithm for fair comparison. Our simulation studies show that the mPI design values safety over efficacy and outperforms other designs in both dose selection and subject assignment. Software to implement the mPI design is freely available. We demonstrate these advantages by implementing the design to find the biologically optimal dose of svldTBI for an immune tolerance conditioning regimen that is applied to achieve mixed chimerism in patients receiving an organ graft and can also implement to phase I cancer trials.

A Unified Decision Framework for Phase I Dose-Finding Designs

Yunshan Duan (UT Austin), Shijie Yuan (Cytel Inc), Yuan Ji (University of Chicago), Peter Mueller (UT Austin)*

The purpose of a phase I dose-finding clinical trial is to investigate the toxicity profiles of various doses for a new drug and identify the maximum tolerated dose. Over the past three decades, various dose-finding designs have been proposed and discussed, including conventional model-based designs, new model-based designs using toxicity probability intervals, and rule-based designs. We present a simple decision framework that can generate several popular designs as special cases. We show that these designs share common elements under the framework, such as the same likelihood function, the use of loss functions, and the nature of the optimal decisions as Bayes rules. They differ mostly in the choice of the prior distributions. We present theoretical results on the decision framework and its link to specific and popular designs like mTPI, BOIN, and CRM. These results provide useful insights into the designs and their underlying assumptions, and convey information to help practitioners select an appropriate design.

Contributed Sessions

Contributed 1

Doubly robust criterion for marginal structural models

Takamichi Baba (The Graduate University for Advanced Studies/Shionogi & Co., Ltd.)*

The semiparametric estimation approach, which includes inverse-probability-weighted and doubly robust estimation using propensity scores, is a standard tool for marginal structural models used in causal inference, and it is rapidly being extended in various directions. On the other hand, although model selection is indispensable in statistical analysis, an information criterion for selecting an appropriate marginal structure has just started to be developed. In this presentation, we derive an Akaike information type of criterion on the basis of the original definition of the information criterion. Here, we define a risk function based on the Kullback-Leibler divergence as the cornerstone of the information criterion and treat a general causal inference model that is not necessarily a linear one. In light of the fact that this field attaches importance to doubly robust estimation, which allows either the model of the assignment variable or the model of the outcome variable to be wrong, we make the information criterion itself doubly robust so that either one can be wrong and it will still be a mathematically valid criterion.

SURE theory in propensity score analysis

Yoshiyuki Ninomiya (The Institute of Statistical Mathematics)*

In the long-term hot methods of propensity score analysis and sparse estimation, we develop the information criterion, a basic tool of statistical analysis, for determining the regularization parameters needed in variable selection. For Gaussian distribution-based causal inference models, we extend Stein's unbiased risk estimation (SURE) theory, which leads to a generalized Cp criterion that has almost no weakness in conventional sparse estimation, and derive an inverse-probability-weighted sparse estimation version of the criterion without resorting to asymptotics. Numerical experiments compare the proposed criterion with the existing criterion derived from a formal argument, and verify that the proposed criterion is superior in almost all cases, that the difference is not negligible in many cases, and that the results of variable selection differ significantly. Real data analysis confirms that the difference between variable selection and estimation by these criteria is actually large.

Robust univariate network meta-analysis without knowledge of within-study correlations: application to primary open-angle glaucoma

Yu-Lun Liu (University of Texas Southwestern Medical Center), Bingyu Zhang (University of Pennsylvania), Yong Chen (University of Pennsylvania)*

Network meta-analysis is an extension of conventional pairwise meta-analysis, and allows the synthesis of the relative effects from multiple interventions in a single analysis. Despite the increasing popularity of network meta-analysis in the field of comparative effectiveness research, it comes with additional limitations and challenges, especially with respect to the reporting bias and the unknown within-study correlations. Specifically, the within-study correlations among treatment comparisons are rarely reported in published literatures. Results indicate that ignoring such correlations when evaluating the pooled treatment effects can be misleading. To deal with this issue, we propose a composite likelihood-based approach that only need to specify marginal model, while allowing the unavailable within-study correlations among treatment comparisons. This proposed method can yield valid statistical inference, even in the setting of that the model is misspecified for correlations. This approach can substantially reduce computational burden. Our approach is motivated by and applied to a network meta-analysis of primary open-angle glaucoma.

Approximate maximum likelihood estimation of the mixture cure model from aggregated data

John Rice (Colorado School of Public Health), Thomas Siebelink (Colorado School of Public Health)*

Research into vaccine hesitancy is a critical component of the public health enterprise, as rates of communicable diseases that are preventable by routine childhood immunization have been increasing in recent years. It is therefore important to estimate proportions of ‚Äúnever-vaccinators,Äù in various subgroups of the population in order to successfully target interventions to improve childhood vaccination rates. However, it is sometimes difficult to obtain individual patient data (IPD) to perform the appropriate time-to-event analyses: state-level immunization information services may only be willing to share aggregated data with researchers. We propose statistical methodology for the analysis of aggregated survival data that can accommodate a cured fraction based on an approximation of the mixture cure model log-likelihood function relying only on summary statistics. We also propose a method to aid in deciding whether the approximation will be adequate, based on Kaplan-Meier estimates of the censoring distribution. We expect these methods to be applicable when there is interest in fitting cure models but where data privacy issues prevent sharing of IPD with researchers.

A Weighted Jackknife Approach Utilizing Linear Model-Based Estimators for Clustered Data

Ruofei Du (University of Arkansas for Medical Sciences), Ye jin Choi (Ohio State University), Ji-Hyun Lee (University of Florida), Songthip Ounpraseuth (University of Arkansas for Medical Sciences), Zhuopei Hu (University of Arkansas for Medical Sciences)*

For analyzing cluster randomized trials (CRTs), standard statistical analysis methods (e.g. linear mixed-effects modeling) require the assumption that the number of clusters is sufficiently large. However, such assumption often cannot be satisfied for real-world studies. Additionally, it is common to observe heterogeneity across the clusters, which leads to less stable statistical inference by a routinely applied method. The small number of clusters combined with cluster level heterogeneity poses a great challenge for the data analysis. We propose a weighted delete-one-cluster Jackknife-based framework utilizing ordinary least squares estimator or generalized least squares estimator to address the concerns. In this scheme, a cluster affected by the heterogeneity is weighted appropriately in estimating the liner combination of the outcome means of study conditions. Compared to some existing representative methods, our simulation studies demonstrated the proposed framework has good operating characteristics with respect to the mean squared error of the estimates and statistical power in testing the difference of the outcome means of conditions.

Contributed 2

Comparing Confidence Distributions in Meta-Analysis

Brinley N. Zabriskie (Brigham Young University), Travis Andersen (Brigham Young University)*

Confidence distributions (CDs), which provide evidence across all levels of significance, are receiving increasing attention, especially in meta-analysis. Meta-analyses allow independent study results to be combined to produce one overall conclusion and are particularly useful in public health and medicine. Several CD approaches for meta-analysis exist, and their relative performance can be assessed via a simulation study. However, only traditional metrics, such as the average coverage of 95% confidence intervals, have been used to measure performances. Clearly, having to choose a significance level in order to evaluate the performance of a CD, where a main benefit is not having to select a significance level, is less than ideal. Here, we develop new metrics to summarize a CD's performance without the need to pre-specify a significance level.

A Two-Stage Decision Making Approach for Safety Studies

Jessica Kim (FDA), Zhipeng Huang (FDA)*

For safety studies, two types of hypothesis testing are considered: detecting a safety signal and ruling out a safety concern. Under the detecting framework, statistical non-significance is often confused with

the conclusion that there is no safety concern. Such a conclusion could be problematic in the presence of low study power or large variability. Under the ruling out framework, determining clinically meaningful margin and larger study size being required could also be some challenges compared to the detecting framework. To overcome such interpretation issues, we propose a Two-Stage Decision-Making (TSDM) approach for safety studies. The proposed TSDM is an adaptive group sequential design that incorporates both detecting a safety signal and ruling out safety concerns into a single study design to increase the probability of making a definite decision. We assess the proposed TSDM approach by conducting Monte Carlo simulations and investigated properties such as operational type I error rate, overall study power based on analytical approximations, overall probability of making a decision, and required sample sizes.

Using functional principal component analysis (fpca) to quantify "active" versus "inactive" sitting patterns derived from wearable sensors

Rong Zablocki (University of California, San Diego), Sheri Hartman (University of California, San Diego), Lindsay Dillon (University of California, San Diego), Loki Natarajan (University of California, San Diego)*

ActiGraph and ActivPAL are commonly used accelerometers. Typically, the ActiGraph is worn at the waist to measure physical activity (PA) and the ActivPAL is worn on the thigh to measure posture. Applying either device alone limits the assessment of true sedentary behavior. The current study overlays both devices time-matched simultaneously to inspect the pattern and variation of subjects' activity while sitting. We hypothesize that those who register more "active" sitting patterns (i.e., higher levels of ActiGraph accelerations) will exhibit better metabolic health than those with "inactive" sitting patterns. We implement a multilevel fpca on 314 post-menopausal women to summarize activity via vector magnitude (VM) from ActiGraph at minute-level within sitting bouts from ActivPAL. VM captures movement acceleration, thus is a measure of PA. Initial analyses showed that over 90% of the subject level variation in VM is loaded on the first 3 principal components, thus dramatically reducing the dimensions from the original minute-level scale. The component scores quantify individuals' activity during sitting bouts, which may differentially impact health outcomes.

Nearest-Neighbor Geostatistical Models for Non-Gaussian Data

Xiaotian Zheng (University of California Santa Cruz), Athanasios Kottas (University of California Santa Cruz), Bruno Sansó (University of California Santa Cruz)*

We develop a class of nearest-neighbor geostatistical models for non-Gaussian data that provides flexibility and scalability. The model is defined on a directed acyclic graph through a weighted combination of first-order spatially varying conditional densities, for each one of a given number of neighbors. It is then extended to a proper spatial process, referred to as the nearest-neighbor mixture transition distribution process (NNMP). We provide conditions to construct general NNMP models with pre-specified stationary marginal distributions. We also establish lower bounds for the resulting strength of the tail dependence, demonstrating the flexibility of NNMPs to quantify multivariate dependence using a bivariate distribution specification. We formulate a Bayesian hierarchical model, focusing on spatially dependent weights. NNMPs lay out a new computational approach to handling spatial data sets, leveraging a mixture model structure to avoid computational issues that arise from large matrix operations. We illustrate the benefits of the NNMP framework using synthetic data examples as well as analyzing sea surface temperature observations from the Mediterranean Sea.

Latent trait shared-parameter mixed-models for missing ordinal ecological momentary assessment data

John Cursio (The University of Chicago)*

A shared parameter model for longitudinal ordinal data collected by ecological momentary assessment (EMA) with missing responses is presented. Mood outcomes collected using EMA were reported by high school students over a period of one week. In this approach, a latent trait representing the responsiveness of subjects is estimated in an item response theory (IRT) sub-model and used as a

covariate in a sub-model for bivariate ordinal mood outcomes. Both likelihood-based and Bayesian-based estimation are shown using statistical software. In the full shared-parameter model, the latent trait of responsiveness is a significant predictor of two mood outcomes, with higher levels of responsiveness associated with improved moods. The full model offers an advantage over missing at random approaches previously used with longitudinal ordinal data containing missing outcomes.

Contributed 3

Modeling Sparse Data Using MLE with Applications to Microbiome Data

Dr. Hani Aldirawi (California State University San Bernardino), Dr. Jie Yang (University of Illinois at Chicago)*

Modeling sparse data such as microbiome and transcriptomics (RNA-seq) data is very challenging due to the exceeded number of zeros and skewness of the distribution. Many probabilistic models have been used for modeling sparse data, including Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models. One way to identify the most appropriate probabilistic models for zero-inflated or hurdle models is based on the p-value of the Kolmogorov-Smirnov test. The main challenge for identifying the probabilistic model is that the model parameters are typically unknown in practice. We derive the maximum likelihood estimator for a general class of zero-inflated and hurdle models. We also derive the corresponding Fisher information matrices for exploring the estimator's asymptotic properties. We include new probabilistic models such as zero-inflated beta binomial and zero-inflated beta negative binomial models. Our application to microbiome data shows that our new models are more appropriate for modeling microbiome data than commonly used models in the literature.

Caution When Inferring the Effect Direction in the Presence of Pleiotropy

Sharon Lutz (Harvard Medical School)*

In genetic association studies, Mendelian Randomization (MR) has gained in popularity as a concept to assess the causal relationship between two phenotypes. The MR Steiger approach has been proposed as a tool that can infer the causal direction between two phenotypes. Through simulation studies, we extend previous work to examine the ability of the MR Steiger approach to correctly determine the effect direction in the presence of pleiotropy, measurement error, and unmeasured confounding. In addition, we examined the performance of these approaches when there is a longitudinal causal relationship between the two phenotypes, under weak instrument variables, and differing distributions for the phenotypes (binary, Poisson, etc). We also applied the MR Steiger method to the COPDGene study, a case-control study of Chronic Obstructive Pulmonary Disease (COPD) in current and former smokers, to examine the role of smoking on lung function in the presence of pleiotropy and measurement error.

Large scale single-cell multi-sample multi-condition data integration using scMerge2

Yingxin Lin (The University of Sydney), Adam Chan (The University of Sydney), Ellis Patrick (The University of Sydney), Jean Yang (The University of Sydney)*

Single-cell profiling such as single-cell RNA-seq have enabled unprecedented insight into the cell identity and functions. The recent emergence of multi-condition multi-sample single-cell large cohort studies allows researchers to investigate different cell states from the same cell type. Effective integration of multiple large cohort studies promises biological insights of cells under different conditions that can not be uncovered with individual study. Here, we present scMerge2, a scalable algorithm that allows data integration of large-scale multi-sample multi-condition studies for various single-cell technologies. Leveraging pseudo-bulk for factor analysis of stably expressed genes and pseudoreplicates, scMerge2 is able to integrate ~200k cells and ~10k genes within an hour. Using a large COVID-19 scRNA-seq data collection with >3 millions cells from >1k samples of ~20 studies globally, we demonstrate that scMerge2 enables large cohort data integration and reveals distinct cell states of COVID-19 patients with

varying degrees of severity. We further illustrate that the effective data integration enables identification of potential signatures that discriminate patient severity.

Statistical Analysis of Metagenomic Hi-C data: Bias Removal and Microbial Genome Retrieval

Yuxuan Du (University of Southern California), Fengzhu Sun (University of Southern California)*

High-throughput chromosome conformation capture (Hi-C) has recently been applied to natural microbial communities and revealed the great potential to simultaneously study multiple genomes and probe active virus-host interactions. Several extraneous factors may influence chromosomal contacts rendering the normalization of Hi-C contact maps essential for downstream analyses. In this presentation, I will report two types of biases in metagenomic Hi-C experiments and introduce a parametric model based on the zero-inflated negative binomial distribution to correct both types of biases and remove spurious interspecies contacts. Normalized metagenomic Hi-C contact maps not only result in lower biases and higher capability to detect spurious contacts, but also substantially benefit the retrieval of high-quality metagenome-assembled genomes (MAGs) from complex microbial ecosystems.

Spatiotemproal Zero-Inflated Bayesian Negative Binomial Regression Using Nearest Neighbor Gaussian Process and Polya-Gamma Mixtures

Qing He (University of Central Florida), Hsin-Hsiung Huang (University of Central Florida)*

Spatialtemporal data analysis with massive zeros are widely used in many areas such as epidemiology and public health. We use a Bayesian framework to fit zero-inflated negative binomial models which introduces a set of latent variables from Polya-Gamma distributions. The posterior Markov chain Monte Carlo algorithm is efficient through Gibbs sampling. The proposed model accommodates varying spatial and temporal random effects through Gaussian process, which gives a flexible covariance structure for the random effects but has computation challenges when the dataset is large. To conquer the computation bottleneck in a Gaussian process, we adopt the nearest-neighbor Gaussian process which approximates the kernel matrix using local experts. For the simulation study, we adopt multiple settings with varying sizes of spatial locations to evaluate the performance of the proposed model such as spatial and temporal random effects estimation and compare the result to other methods. We also apply the proposed algorithm to the COVID-19 data to examine death rates among counties with high and low social vulnerability levels in Florida, USA.

Contributed 4

Time Series Forecasting and Forecast Intervals with Random Forests

Barbara Bailey (San Diego State University)*

Random forests consist of an ensemble of decision trees for regression and have successfully been used for prediction in wide range of applications. The modeling, forecasting, and construction of forecast intervals for time series data are investigated. The quantile random forest is used in the construction of forecast intervals. Results are compared to the stationary bootstrap to generate realizations of the time series to be used in the building of each tree in the random forest and in the construction of forecast intervals. Applications to financial data are presented.

Flexible variable selection in the presence of missing data

Brian D Williamson (Kaiser Permanente Washington Health Research Institute and Fred Hutchinson Cancer Research Center), Ying Huang (Fred Hutchinson Cancer Research Center and University of Washington)*

In many applications, it is of interest to identify a parsimonious set of features, or panel, from multiple candidates that achieves a desired level of performance in predicting a response. This task is often complicated in practice by missing data arising from the sampling design or other random mechanisms. Most recent work on variable selection in missing data contexts relies in some part on a finite-dimensional statistical model (e.g., a generalized or penalized linear model). In cases where this model is

misspecified, the selected variables may not all be truly scientifically relevant and can result in panels with suboptimal classification performance. To address this limitation, we propose several nonparametric variable selection algorithms combined with multiple imputation to develop flexible panels in the presence of missing-at-random data. We outline strategies based on the proposed algorithms that achieve control of commonly used error rates. Through simulations, we show that our proposals have good operating characteristics and result in panels with higher classification performance compared to several existing penalized regression approaches.

Sampling from multimodal distributions using tempered Hamiltonian transitions

Joonha Park (University of Kansas)*

Hamiltonian Monte Carlo (HMC) methods are widely used to draw samples from unnormalized target densities due to high efficiency and favorable scalability with respect to increasing space dimensions. However, HMC struggles when the target distribution is multimodal, because the maximum increase in the potential energy function along the simulated path is bounded by the initial kinetic energy. In this paper, we develop a Hamiltonian Monte Carlo method where the constructed paths can travel across high potential energy barriers. This approach enables frequent jumps between the isolated modes of the target density. Compared to other tempering methods, our method has a distinctive advantage in the Gibbs sampler settings, where the target distribution changes at each step. We develop a practical tuning strategy and demonstrate that it can construct globally mixing Markov chains targeting high-dimensional, multimodal distributions, using mixtures of normals and a sensor network localization problem.

Model-agnostic Explanations of Survival Prediction Models

Krithika Suresh (University of Colorado), Carsten Görg (University of Colorado), Karen Kanaster (University of Colorado), Debashis Ghosh (University of Colorado)*

Advanced machine learning methods, capable of capturing complex and non-linear relationships, can be used in biomedical research to accurately predict time-to-event outcomes. However, these methods have been criticized as “black boxes” that are not interpretable and thus are difficult to trust in making important clinical decisions. Explainable machine learning proposes the use of model-agnostic explainers that can be applied to predictions from any complex model. These explainers describe which patient’s characteristics are contributing to their prediction, and thus provide insight into how the model arrived at that prediction. The application of explainers to survival prediction models can provide explanations for an individual’s overall survival curve as well as survival predictions at particular follow-up times. Here, we present an approach for obtaining these explanations from any survival prediction model. We extend the local interpretable model-agnostic explainer (LIME) for classification outcomes to be applied to survival prediction models using a landmarking and multi-task learning framework. We illustrate application of the new methodology using a real-world data set.

Parallel Tempering With a Variational Reference

Nikola Surjanovic (University of British Columbia), Saifuddin Syed (University of British Columbia), Alexandre Bouchard-Côté (University of British Columbia), Trevor Campbell (University of British Columbia)*

Sampling from multi-modal and high-dimensional target distributions is a challenging task that is often required to perform Bayesian inference. Parallel tempering (PT) methods address this problem by constructing a Markov chain on an expanded state space that simultaneously samples from a sequence of distributions lying on an annealing path from the prior to the target. In this work we consider generalized annealing paths that start from a variational reference. The reference distribution is tuned to minimize the forward KL divergence to the target distribution using a simple (gradient-free) moment-matching procedure. We apply the method to several posterior inference problems, finding that PT with a variational reference can greatly improve performance. The proposed methodology is particularly

useful in the typical case where the prior and posterior are almost mutually singular and the geometry of the posterior is complex.

Contributed 5

Confidence Envelopes for Model Selection Criteria and Post-Model Selection Inference

A.Alexandre Trindade (Texas Tech University)*

In choosing a candidate model in likelihood-based modeling via an information criterion, the practitioner must decide how far up the ranked list to look. Motivated by this necessity, we construct an uncertainty band for a generalized information criterion (GIC), defined as a criterion for which the limit in probability is identical to that of the normalized log-likelihood. This includes common special cases such as AIC & BIC. The method starts from the joint asymptotic normality of the GIC values, and proceeds by deriving the (asymptotically) exact distribution of the minimum. Inversion of this CDF then provides the desired quantiles. The joint asymptotics of the GICs is derived in three cases of classical interest: IID, regression, time series. The methodology's performance is assessed on simulated data via coverage probabilities of nominal upper quantiles, and compared to the bootstrap. Both methods give coverages close to nominal for large samples, but the bootstrap is two orders of magnitude slower. The methodology's ability to produce confidence intervals for individual parameters by pivoting the CDF of the minimum GIC, naturally accounts for model selection uncertainty.

Information criteria for detecting change-points in the Cox proportional hazards model

Ryoto Ozaki (The Graduate University for Advanced Studies/Chugai Pharmaceutical Co., Ltd.)*

The Cox model assumes proportional hazards, but it has been shown that the assumption does not hold in cases such as the delayed onset of a treatment effect. In such a situation, the survival curves are expected to overlap for a certain period after the start of treatment, and then the difference between the curves increases. As this is considered to be an acute change in the hazard ratio function, change-point analysis is important in survival time analysis. Hence, this paper considers the Cox proportional hazards model with change-points and derives AIC-type information criteria for detecting those change-points. Accordingly, we construct specific asymptotic theories in that model by using the partial likelihood estimation method. By applying the original AIC derivation method, we propose information criteria that are with penalties for change-points much larger than twice the number of that parameters and are mathematically guaranteed. Numerical experiments confirm that the proposed criterion, in comparison to a formal AIC that penalizes twice the number of parameters, more accurately approximates the asymptotic bias to the risk function.

Real-Time Change Point Detection in High-Dimensional Linear Models

Suthakaran Ratnasingam (California State University, San Bernardino), Wei Ning (Bowling Green State University)*

In this article, we propose a procedure to monitor the structural changes in the penalized regression model for high-dimensional data sequentially. Our approach utilizes a given historical data set to perform both variable selection and estimation simultaneously. The asymptotic properties of the test statistics are established under the null and alternative hypotheses. The finite sample behavior of the monitoring procedure is investigated with simulation studies. The proposed method is applied to a real data set to illustrate the detection procedure.

Clustering in High Dimensions

William Lippitt (University of Colorado Denver AMC), Nichole E Carlson (University of Colorado Denver AMC), Jaron Arbet (UCLA), Tasha Fingerlin (University of Colorado Denver AMC; National Jewish Health), Lisa Maier (University of Colorado Denver AMC; National Jewish Health), Katerina Kechris (University of Colorado Denver AMC)*

With large volumes of clinical, imaging, and genomic/genetic data available on patients, a common question for complex diseases has become whether new disease subtypes can be identified by analyzing data heterogeneity in an unsupervised fashion. Often, the data are highly correlated, redundant, and noisy. The performance of standard Gaussian Mixture Model (GMM) and GMM-based clustering approaches in this setting is unclear. The purpose of this work was to investigate the performance of GMMs in noisy, highly correlated, moderate-to-high dimension contexts. In particular, we discovered poor performance is often driven by the variance-as-relevance assumption. This is the same assumption made in standard dimension reduction in PCA, namely that high variance features/PCs are “relevant” and should be retained or more finely modelled while lower variance features/PCs are “irrelevant” and should be discarded or more coarsely modelled. A crude discriminative dimension reduction technique, inspired by studies of gene expression data and built from PCA and the Shapiro-Wilk normality test, was introduced to demonstrate how this assumption might be addressed.

Modeling Dynamic Correlation Structure in Multi-Omics Data from Single-Cell Experiments

Zichen Ma (Clemson University), Zhen Yang (University of Southern California), Yen-Yi Ho (University of Southern California)*

With the recent advance in technologies to profile multi-omics data at the single-cell level, integrative multi-omics data analysis has been increasingly popular. Information such as methylation changes, chromatin accessibility, and gene expression are often jointly collected in a single-cell experiment. In biomedical studies, it is of interest to study the associations between various data types and to examine how these associations might change according to other factors, such as cellular states, or other data types. However, since each data type usually has a distinct marginal distribution, joint analysis of these changes of associations using multi-omics data is statistically challenging. This talk will introduce flexible Bayesian modeling frameworks for studying dynamics correlation structures between gene expression read counts and other data types. The performance of the proposed frameworks will be demonstrated through a series of simulation studies. We will also illustrate the implementation of our proposed modeling frameworks using datasets from two single-cell studies: the role of BRAF inhibitor resistance in melanoma patients and a study of mouse gastrulation.

Student Paper Sessions

Student Paper 1

A Double Robust Estimator For Mann Whitney Wilcoxon Rank Sum Test When Applied for Causal Inference in Observational Studies

Ruohui Chen (University of California, San Diego)*

The Mann-Whitney-Wilcoxon MWW rank sum test (MWWRST) is widely used to compare two treatment groups in randomized control trials when data distributions are highly skewed, especially in the presence of outliers. As the MWWRST generally yields invalid inference when applied to observational study data, Wu et al. (2014) introduced an approach to address confounding effects by incorporating the inverse probability weighting technique into this rank-based statistic. More importantly, their approach addressed limitations of an earlier attempt by Rosenbaum (2002) based on a randomization inference technique by positing a constant treatment effect between two potential outcomes across all subjects. This assumption not only completely ignores subject level differences, but also is unverifiable in real studies. In this paper, we address an important limitation in Wu et al. (2014) by extending their approach to a doubly robust setting to provide more robust causal inference by integrating functional response models with the the inverse probability weighting and mean score imputation. We demonstrate performances of the approach through both simulated and real study data.

Model-assisted analyses of cluster-randomized experiments

Fangzhou Su (UC Berkeley)*

Cluster-randomized experiments are widely used due to their logistical convenience and policy relevance. To analyze them properly, we must address the fact that the treatment is assigned at the cluster level instead of the individual level. Standard analytic strategies are regressions based on individual data, cluster averages, and cluster totals. These methods are often motivated by models with strong and unverifiable assumptions, and the choice among them can be subjective. Without any outcome modeling assumption, we evaluate these regression estimators and the associated robust standard errors from a design-based perspective where only the treatment assignment itself is random and controlled by the experimenter. We demonstrate that regression based on cluster averages targets a weighted average treatment effect, regression based on individual data is suboptimal in terms of efficiency, and regression based on cluster totals is consistent and more efficient with a large number of clusters. We highlight the critical role of covariates in improving estimation efficiency, and illustrate the efficiency gain via both simulation studies and data analysis.

Fighting Noise with Noise: Causal Inference with Many Candidate Instruments

Xinyi Zhang (University of Toronto), Linbo Wang (University of Toronto), Stanislav Volgushev (University of Toronto), Dehan Kong (University of Toronto)*

Instrumental variable methods provide useful tools for inferring causal effects in the presence of unmeasured confounding. To apply these methods with large-scale data sets, a major challenge is to find valid instruments from a possibly large candidate set. In practice, most of the candidate instruments are often not relevant for studying a particular exposure of interest. Moreover, not all relevant candidate instruments are valid as they may directly influence the outcome of interest. In this article, we propose a data-driven method for causal inference with many candidate instruments that addresses these two challenges simultaneously. A key component of our proposal is a novel resampling method that constructs pseudo variables to remove irrelevant candidate instruments having spurious correlations with the exposure. Theoretical and synthetic data analyses show that the proposed method performs favourably compared to existing methods. We apply our method to a Mendelian randomization study estimating the effect of obesity on health-related quality of life.

Estimation based on nearest neighbor matching: from density ratio to average treatment effect

Zhexiao Lin (University of Washington), Peng Ding (University of California, Berkeley), Fang Han (University of Washington)*

Nearest neighbor (NN) matching is a conceptually natural and practically well-used tool to align data sampled from different groups. In a landmark paper, Abadie and Imbens (2006) provided the first large-sample analysis of NN matching. Their theory, however, requires a crucial assumption that the number of NNs, M, is fixed. We reveal something new out of their study and show that, once allowing M to diverge with the sample size, an intrinsic statistic in their analysis actually constitutes a consistent estimator of the density ratio. Furthermore, we show that through selecting a suitable M, this statistic can attain the minimax lower bound of estimation over a Lipschitz density function class. Consequently, with a diverging M, the NN matching with Abadie and Imbens (2011),Äôs bias correction provably yields a doubly robust estimator of the average treatment effect and is semiparametrically efficient if the density functions are sufficiently smooth and the outcome model is appropriately specified. It can thus be viewed as a precursor of the recently proposed double machine learning estimators.

Student Paper 2

Inference after latent variable estimation for single-cell RNA sequencing data

Anna Neufeld (University of Washington)*

In the analysis of single-cell RNA sequencing data, researchers often first estimate a latent variable that characterizes variation between cells, and then test each gene for association with the estimated latent variable. If the same data are used for both of these steps, then standard methods for computing p-values and confidence intervals in the second step will fail to achieve standard statistical guarantees such as Type 1 error control or nominal coverage. Furthermore, approaches such as sample splitting that can be fruitfully applied to solve similar problems in other settings are not applicable in this context. We introduce count splitting, an extremely flexible framework that allows us to carry out valid inference in this setting, for virtually any latent variable estimation technique and inference approach, under a Poisson assumption. We demonstrate the Type 1 error control and power of count splitting in a simulation study, and apply count splitting to a dataset of pluripotent stem cells differentiating to cardiomyocytes.

An Integrated Bayesian Framework for Multi-omics Prediction and Classification

Anupreet Porwal (Department of Statistics, University of Washington Seattle, WA 98195, USA), Himel Mallick (Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, New Jersey 07065, U.S.A.), Satabdi Saha (Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, USA), Vladimir Svetnik (Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, New Jersey 07065, U.S.A.), Erina Paul (Biostatistics and Research Decision Sciences, Merck & Co., Inc., Rahway, New Jersey 07065, U.S.A.)*

With the growing commonality of multi-omics datasets, there is now increasing evidence that integrated omics profiles lead to more efficient discovery of clinically actionable biomarkers that enable better disease outcome prediction and patient stratification. Several methods exist to perform host phenotype prediction from cross-sectional, single-omics data modalities but decentralized frameworks that jointly analyze multiple time-dependent omics data to highlight the integrative and dynamic impact of repeatedly measured biomarkers is limited. We propose a novel Bayesian ensemble method to consolidate prediction by borrowing information across several longitudinal and cross-sectional omics data layers. Unlike existing paradigms, our approach enables uncertainty quantification in prediction as well as interval estimation for a variety of quantities of interest. We apply our method to four published multi-omics datasets and demonstrate that it recapitulates known biology in addition to providing novel insights while also outperforming existing methods in predictive performance. Our open-source software is publicly available at https://github.com/himelmallick/IntegratedLearner.

A multivariate approach to conditional independence testing in microbial network construction

Hongjiao Liu (University of Washington), Yunhua Xiang (University of Washington), Michael C. Wu (Fred Hutchinson Cancer Research Center)*

Graphical modeling helps elucidate the complex interrelationships among a set of features. A key step in graphical modeling is to assess the conditional dependence between features. While common approaches for evaluating conditional dependence treat individual features as univariate variables, a multivariate approach could be advantageous in certain situations, e.g., when features are each composed of multiple sub-features. In particular, for a pair of features, if there are heterogeneous relationships present among their sub-features, an aggregated univariate approach might result in a loss of statistical power. Here we propose a flexible and nonparametric multivariate testing framework, Conditional RV, for assessing the conditional dependence between two multivariate features in a graphical model. We demonstrate the performance of Conditional RV in the context of microbial association network construction, using both simulation studies and real data application. In the presence of heterogeneous relationships among sub-features, we show that Conditional RV has an improved power in detecting conditional dependence compared to univariate competing methods.

IBD-based estimation of X chromosome effective population size with application to sex-specific demographic history

Ruoyi Cai (Department of Biostatistics, University of Washington), Brian Browning (Department of Biostatistics and Department of Medicine, University of Washington), Sharon Browning (Department of Biostatistics, University of Washington)*

A powerful approach to estimate effective population size (Ne) in the recent past is to reconstruct coalescence history through identity-by-decent (IBD) segments. While a number of methods have been developed for estimating Ne from IBD segments carried in autosomes, no such effort has been made with X chromosome IBD segments. In this work, we proposed an IBD-based method to estimate X chromosome Ne through probabilistic modeling of the X chromosome coalescence process. We also showed that the comparison of X chromosome Ne to autosome Ne provides information on sex-specific effective population sizes. Autosome and X chromosome Ne estimated using our method can recover the correct Ne in simulated populations. In analysis of real-world populations using data from UK Biobank and TOPMed project, we find results to be consistent with equal male and female effective population sizes during the recent past in the UK white British and UK Indian populations and in an African American population. This study emphasized that analysis of the X chromosome could enable us to understand sex-specific population dynamics that may not be revealed by the analysis of autosomes alone.

Integrative cross-omics and cross-context analysis elucidates molecular links underlying complex diseases and traits

Yihao Lu (Department of Public Health Sciences, The University of Chicago), Meritxell Oliva (Department of Public Health Sciences, The University of Chicago), Brandon L. Pierce (Department of Public Health Sciences, The University of Chicago), Jin Liu (Centre for Quantitative Medicine, Program in Health Services & Systems Research, Duke-NUS Medical School), Lin S. Chen (Department of Public Health Sciences, The University of Chicago)*

The integration of association results from genome-wide association studies with large-scale genomic, particularly multi-omic, studies provides opportunities and challenges to elucidate the disease/trait mechanisms and their operating cellular contexts. To integrate multiple sets of genetic/genomic association statistics each from multiple contexts, we propose a method, X- ING (Cross-INtegrative Genomics), that enables cross-omics and cross-context integration of association summary statistics. X-ING implements a hierarchical Bayesian model that estimates the latent binary association status of each statistic, and accounts for the major patterns shared across omics traits and contexts to improve power and precision. Broadly, X-ING enables the cross-feature integration of effects from multivariate

contexts/studies. We apply X-ING in a multi-tissue multi-omics analysis using genetics, methylome and transcriptome data from the Genotype-Tissue Expression project. Our integrative multi-tissue expression and methylation quantitative-trait-locus (QTL) analysis examines the roles of cis- and trans-QTLs in the etiologies of complex diseases/traits and the e/mQTL effect sharing patterns.

Student Paper 3

Estimating relative survival after TEVAR for thoracic aortic aneurysm

Hang Nguyen (Southern Methodist University, Dallas, TX), Haekyung Jeon-Slaughter (VA North Texas Health Care System, Dallas, TX), Daniel F. Heitjan (Southern Methodist University, Dallas, TX; UT Southwestern Medical Center, Dallas, TX)*

Many registries exist to preserve data on patients treated with surgical devices. Using registry data presents a statistical challenge due to right censoring and left-truncation. A popular estimand is the relative survival, or the ratio of survival functions in the registry patients to that of a control population. For large sample sizes, one can compute relative survival directly from the Kaplan-Meier curve in the disease group and the control life table. For modest sample sizes, one can estimate survival curves for the disease group by modeling the dependence of the mortality hazard on truncation time and time since surgery. We use a flexible logistic discrete-time model to estimate the survival hazard in TEVAR patients as a function of age at TEVAR and time since surgery, to estimate the relative survival for these patients to the general US population, matched on age, race, and sex. The relative survival appears to follow a U-shaped curve, exceeding 1 for younger and older patients and less than 1 for middle group. Surprisingly, relative survival is highest in the very old. Whether this is a real effect or a consequence of an unknown sampling bias is a topic for further study.

Impact of correlations between prioritized outcomes on the net benefit and its estimate by generalized pairwise comparisons

Kanako Fuyama (Hokkaido University), Mitsunori Ogawa (The University of Tokyo), Junki Mizusawa (National Cancer Center), Koji Oba (The University of Tokyo)*

The net benefit based on prioritized outcomes is an emerging benefit-risk metric in clinical trials. Although previous research has demonstrated that the correlations between outcomes impact the net benefit and its estimate, the direction and magnitude of this impact remain unclear. In this study, we investigated the impact of correlations between two binary or Gaussian variables on the true net benefit values via theoretical consideration and numerical computation. We also explored the impact of correlations between survival and categorical variables on the net benefit estimates in the presence of right censoring by simulation and application to clinical trial data. Our investigation revealed that the true net benefit values were impacted by the correlations in various directions depending on the outcome distributions, though this direction was governed by a simple rule in the case with binary endpoints. Our simulation also showed that their estimates could be severely biased by right censoring, and that the direction and magnitude of this bias were associated with the correlations. The impact of correlations should be considered when interpreting the net benefit and its estimate.

Privacy-Preserving and Communication-Efficient Causal Inference for Hospital Quality Measurement

Larry Han (Harvard University), Yige Li (Harvard University), Bijan Niknam (Harvard University), Jose Zubizarreta (Harvard University)*

Data sharing can improve hospital quality measurement, but sharing patient-level data between hospitals is often infeasible. Motivated by the evaluation of Cardiac Centers of Excellence (CCE), we propose a method to safely leverage information from peer hospitals to improve the precision of quality estimates. We develop a doubly robust estimator that is privacy-preserving (requiring only summary statistics be shared) and communication-efficient (requiring only one round of communication). We contribute to the quality measurement and causal inference literatures by developing a framework to

assess treatment-specific performance in hospitals. We propose a penalized regression approach on summary statistics of the influence functions for efficient estimation and valid inference. The proposed estimator is data-adaptive, downweighting hospitals with different case-mixes from the target hospital for bias reduction and upweighting hospitals with similar case-mixes for efficiency gain. We find that our estimator improves precision of treatment effect estimates by 34% to 86% for target hospitals, qualitatively altering the evaluation of treatment effects in 22 of 51 hospitals.

Wedding Table Cluster Prior with Exchangeability and Parallel Processing

Charles Harrison (University of Central Florida)*

We consider choosing the mass parameter for a new part in a prior distribution on partitions. We discuss a variety of properties including exchangeability, parallel processing, and the expected number of clusters. Next, we apply our prior for clustering vector-variate and matrix-variate data using a Gaussian likelihood function.

Student Paper 4

A computationally efficient approach to fitting large scale penalized cox models

Aliasghar Tarkhan (PhD student, Department of Biostatistics, University of Washington), Noah Simon (Associate Prof., Department of Biostatistics, University of Washington)*

In many biomedical applications, we measure the outcome as a ‚Äútime-to-event‚Äù(e.g., disease progression or death). Cox regression is a widely used tool to assess the connection between a patient‚Äôs characteristics and this outcome. With advancements in data acquisition technologies, it is increasingly common to collect extremely large datasets. Standard approaches (such as the lasso method) to fitting the Cox proportional hazards regression model perform well via variable selection and estimation but tend to fail for large-scale or ultra-high dimensional datasets due to computational instability and/or memory limits. To address this, we propose a modification to the partial likelihood that facilitates fitting the cox model on larger datasets and enables the use of stochastic optimization techniques. In particular, our proposed framework enables data to be read off the hard drive in chunks to update our model sequentially. We apply stochastic proximal gradient descent in our framework to fit Cox regression models with the elastic net penalty. We show that our proposed framework performs well when the number of observations n and features p grows (particularly with $p >> n$).

Multilevel hybrid principal components analysis for region-referenced functional EEG data

Emilie Campos O'Banion (University of California, Los Angeles), Aaron Wolfe Scheffler (University of California, San Francisco), Donatello Telesca (University of California, Los Angeles), Catherine Sugar (University of California, Los Angeles), Damla Senturk (University of California, Los Angeles)*

Electroencephalography experiments produce region-referenced functional data representing brain signals in the time or the frequency domain collected across the scalp. The data typically also have a multilevel structure with high-dimensional observations collected across multiple experimental conditions or visits. Common analysis approaches reduce the data complexity by collapsing the functional and regional dimensions, where event-related potential (ERP) features or band power are targeted in a pre-specified scalp region. This practice can fail to portray more comprehensive differences in the entire ERP signal or the power spectral density (PSD) across the scalp. Building on the weak separability of the high-dimensional covariance process, the proposed multilevel hybrid principal components analysis (M-HPCA) utilizes dimension reduction tools from both vector and functional principal components analysis to decompose the total variation into between- and within-subject variance. The resulting model components are estimated in a mixed effects modeling framework via a computationally efficient minorization-maximization algorithm.

Bayesian Functional Partial Membership Models

Nicholas Marco (University of California, Los Angeles), Damla Senturk (University of California, Los Angeles), Donatello Telesca (University of California, Los Angeles)*

Partial membership models, or mixed membership models, are a flexible unsupervised clustering method that allows observations to belong to multiple clusters at the same time. In this paper, we propose a Bayesian partial membership model for functional data. By using the multivariate Karhunen-Lo$V^{®}$ve theorem, we are able to derive a scalable model that does not make many assumptions on the covariance structure of the data. Compared to previous work on partial membership models, our proposed model is more flexible and allows for direct interpretation of the mean and covariance structure. To illustrate the usefulness of our model, we fit our partial membership model on EEG signals from a cohort containing children with Autism Spectrum Disorder (ASD). We found that the results from our model tend to agree with the results previously found in the scientific literature, however our model allows clinicians to analyze the data in a novel way.

Outlier Detection for Brain Network Data

Pritam Dey (Department of Statistical Science, Duke University), Zhengwu Zhang (Statistics and Operations Research, University of North Carolina at Chapel Hill), David Dunson (Department of Statistical Science, Duke University)*

It has become routine in neuroscience studies to measure brain networks for different individuals using neuroimaging. These networks are typically expressed as adjacency matrices, with each cell containing a summary of connectivity between a pair of brain regions. There is an emerging statistical literature describing methods for the analysis of such multi-network data. However, there has been essentially no consideration of the important problem of outlier detection. In particular, for certain subjects, the neuroimaging data are so poor quality that the network cannot be reliably reconstructed. For such subjects, the resulting adjacency matrix may be mostly zero or exhibit a bizarre pattern not consistent with a functioning brain. These outlying networks may serve as influential points, contaminating subsequent statistical analyses. We propose a simple method for network outlier detection (NOD) relying on an influence measure under a hierarchical generalized linear model for the adjacency matrices. An efficient computational algorithm is described, and our NOD method is illustrated through simulations and an application to data from the UK Biobank.

Proximal MCMC for Bayesian Inference of Constrained and Regularized Estimation

Xinkai Zhou (Department of Biostatistics, UCLA), Eric Chi (Department of Statistics, Rice University), Hua Zhou (Department of Biostatistics, UCLA)*

This paper advocates proximal Markov Chain Monte Carlo (ProxMCMC) as a generic Bayesian inference framework for constrained or regularized estimation. Originally developed in the Bayesian imaging literature, ProxMCMC deploys the Moreau-Yosida envelop for a smooth approximation of the total variation regularization term, fixes nuisance and regularization parameters as constants, and relies on the Langevin algorithm for the sampling of the posterior. We extend the ProxMCMC to the full Bayesian framework with modeling and data adaptive estimation of all parameters including the regularization parameter. More efficient sampling algorithms such as the Hamiltonian Monte Carlo are employed to scale ProxMCMC to high-dimensional problems. Analogous to the proximal algorithms in optimization, ProxMCMC offers a versatile and modularized procedure for the inference for constrained and non-smooth problems. The power of ProxMCMC is illustrated on various statistical estimation and machine learning tasks. The inference in these problems is traditionally considered difficult from both frequentist and Bayesian perspectives.

Student Paper 5

Shape-Based Clustering of Daily Weigh-In Trajectories using Dynamic Time Warping

Samantha Bothwell (Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado), Alex Kaizer (Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado), Ryan Peterson (Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado), Danielle Ostendorf (Department of Medicine, Division of Endocrinology, Metabolism, and Diabetes, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA.), Victoria Catenacci (Department of Medicine, Division of Endocrinology, Metabolism, and Diabetes, University of Colorado Anschutz Medical Campus, Aurora, Colorado, USA.), Julia Wrobel (Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Anschutz Medical Campus, Aurora, Colorado)*

Smart-scales are a new tool for monitoring weight change and weigh-in behavior. These scales give the opportunity to discover patterns in the frequency that individuals weigh themselves over time, and how these patterns are associated with overall weight loss. Our motivating data come from an 18-month behavioral weight loss study of 55 adults classified as overweight or obese who were instructed to weigh themselves daily. Adherence to daily weigh-in routines produces a binary times series for each subject, indicating whether a participant weighed in on a given day. To characterize weigh-in patterns by shapes rather than overall adherence, we propose using hierarchical clustering with Dynamic Time Warping (DTW) a distance metric optimized for continuous data. We perform an extensive simulation study to evaluate the performance of DTW compared to Euclidean and Jaccard distances to recover underlying patterns in binary time series. In addition, we compare cluster performance using cluster validation indices under different linkages. We apply conclusions from the simulation to cluster our motivating data and summarize observed weigh-in patterns.

Bayesian random change point mixed model analysis of cognitive performance trajectories to identify eligible patients for randomized clinical trials

Lianlian Du (Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA), Rebecca Langhough Koscik (Department of Medicine, University of Wisconsin-Madison, Madison, WI, USA), Tobey J Betthauser (Department of Medicine, University of Wisconsin-Madison, Madison, WI, USA), Sterling C. Johnson (Department of Medicine, University of Wisconsin-Madison, Madison, WI, USA), Bret Larget (Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA), Richard J. Chappell (Department of Biostatistics and Medical Informatics and department of statistics, University of Wisconsin-Madison, Madison, WI, USA)*

Cognitive decline rates in Alzheimer,Äôs disease (AD) and other dementias may accelerate significantly in preclinical phases. Changepoint analysis can reveal when the change starts to diverge from the pattern of normal aging. Wisconsin Registry for Alzheimer,Äôs Prevention (WRAP) participants with >=3 Preclinical Alzheimer,Äôs Cognitive Composite (PACC3) scores, dementia-free at baseline PACC3 was included (n=1068). We proposed a Bayesian random change point mixed model to estimate fixed (group-level) and random (person-level) change points (CPs), slopes pre- and post-CPs, and intercepts at CPs. We examined how these parameters varied across predictors of interest. The CPs of this innovative mixed model are subject to variation. This feature is crucial for real-world evidence. All random effects PACC3 estimates differed by last observed cognitive status, sex/gender, and biomarkers. APOE-e4 random-effects estimates differed by CP and post-CP slopes. Predicting individual cognition trajectories and change points may provide an opportunity for early intervention by finding the high-risk participants at the right time for the clinical trial, and this new methodology is a useful means.

Joint Modeling of Longitudinal Processes and Mortality via Mixed Effects State Space Models with Applications to Dialysis Data

Ya Luo (Novartis), Mingzhao Hu (University of California, Santa Barbara), Yuedong Wang (University of California, Santa Barbara)*

For hemodialysis patients, mortality rates remain high while their quality of life remains low. Based on their strict schedules, there is a large amount of longitudinal data on clinical indicators of which only a small portion has been explored for better personalized treatment. This paper develops a new joint modeling framework to model multivariate longitudinal variables using a mixed effects state space model (MESSM) and death using a logistic regression model. The proposed joint model can handle many different longitudinal profiles and avoids the proportional hazard assumption. These new statistical methods can be used to extract information from individual profiles, explore dynamic relationships within and across longitudinal variables, identify risk factor for mortality, and support individualized clinical decisions based on predictions of future values of longitudinal variables as well as risks of mortality. We propose a new computational method for Kalman filtering, smoothing, and likelihood calculation that scales linearly in the number of subjects. We apply the proposed methods to create an online prediction model for mortality in hemodialysis patients.

Variable Importance for Fixed Effects in Linear Mixed Model

Yongzhe Wang (University of Washington), Lingbo Ye (University of Washington), Zifan Yu (University of Washington)*

Many scientific applications are of interest to evaluate the relative shares of influence of variables in a given model through the change in prediction values or metrics, namely to explore the variable importance for covariates. Researchers have proposed different approaches to investigate variable importance for cross-sectional data with parametric and non-parametric models already. However, this topic is less brought up in the context of longitudinal data. To tackle the problem, we introduced a variable importance measurement (VIM), invented by Lindeman, Merenda, and Gold, for fixed effects in the linear mixed effect model. To cooperate with the nature of cluster effects in longitudinal data, we used marginal and conditional R^2 to obtain the variable importance, which offered two interpretations of the VIM through the improvement of R^2 from the subject level and the population level. Meanwhile, it was robust for assessing contributions to fixed effects under the presence of multicollinearity. Throughout simulations, we showed that our proposed VIM for covariates matched the true rank of covariates in data generation process for simulating longitudinal data.

Nonlinear Mixed-Effects Models for HIV Viral Load Trajectories Before and After Antiretroviral Therapy Interruption, Incorporating Left Censoring

Sihaoyu Gao (University of British Columbia), Lang Wu (University of British Columbia), Tingting Yu (Harvard Pilgrim Health Care Institute and Harvard Medical School), Roger Kouyos (University of Zurich), Huldrych F. Gunthard (University of Zurich), Rui Wang (Harvard T.H. Chan School of Public Health)*

Characterizing features of the viral rebound trajectories and identifying host, virological, and immunological factors that are predictive of the viral rebound trajectories are central to HIV cure research. In this paper, we investigate if key features of HIV viral decay and CD4 trajectories during antiretroviral therapy (ART) are associated with characteristics of HIV viral rebound following ART interruption. Nonlinear mixed effect (NLME) models are used to model viral load trajectories before and following ART interruption, incorporating left censoring due to lower detection limits of viral load assays. A stochastic approximation EM (SAEM) algorithm is used for parameter estimation and inference. To circumvent the computational intensity associated with maximizing the joint likelihood, we propose an easy-to-implement three-step method. We evaluate the performance of this method through simulation studies and apply it to data from the Zurich Primary HIV Infection Study. We find that some key features of viral load during ART (e.g., viral decay rate) are significantly associated with important characteristics of viral rebound following ART interruption (e.g., viral set point).

Student Paper 6

Federated Offline Reinforcement Learning

Doudou Zhou (University of California, Davis)*

Offline reinforcement learning (RL) aims to learn an optimal policy based on a dataset collected a priori. Under the multi-site setting where the datasets are distributed among different sites, such as the mobile health data, the RL algorithms using individual-level data can often not be applied directly due to privacy concerns or communication costs. However, analysis based on multiple sites is necessary when the site-specific covariates are of interest or the data size of a single site is small. To solve such a problem, we propose a federated dynamic treatment regimes algorithm (FDTR), which only requires one communication among the sites by sharing summary statistics. As a result, FDTR is communication-efficient and preserves privacy. We provide theoretical results for FDTR, which guarantee suboptimality for the learned policies and are shown to be better than the local ones. Extensive simulations demonstrate the effectiveness of FDTR. The method is then applied to a sepsis data set in multiple sites to illustrate its use in clinical settings.

Metaheuristics for finding efficient longitudinal designs for sustained release lithium in bipolar disorder

Mitchell Schepps (UCLA Department of Biostatistics, Los Angeles, CA), Weng Kee Wong (UCLA Department of Biostatistics, Los Angeles, CA)*

When there are a few candidate designs for implementation in pharmacometrics, a common method to select the design is to adopt a model-based approach and determine the design with the best value of a pre-selected design criterion among the candidate designs. The design criterion is formulated as a scalar function the Fisher information matrix, which can be challenging to evaluate for non-linear mixed effects models. We propose using nature-inspired metaheuristic algorithms to search for efficient model-based designs with user selected number of time points to optimize the design criterion. We discuss use of metaheuristics as a general purpose optimization tool and apply it to design efficient longitudinal studies for bipolar patients with and without a genetic covariate and treated with lithium.

Estimating Optimal Infinite Horizon Dynamic Treatment Regimes via pT-Learning

Wenzhuo Zhou (University of Illinois Urbana Champaign), Ruoqing Zhu (University of Illinois Urbana Champaign), Annie Qu (University of California Irvine)*

Recent advances in mobile health (mHealth) technology provide an effective way to monitor individuals' health statuses and deliver just-in-time personalized interventions. However, the practical use of mHealth technology raises unique challenges to existing methodologies on learning an optimal dynamic treatment regime. Many mHealth applications involve decision-making with large numbers of intervention options and under an infinite time horizon setting where the number of decision stages diverges to infinity. In addition, temporary medication shortages may cause optimal treatments to be unavailable, while it is unclear what alternatives can be used. To address these challenges, we propose a Proximal Temporal consistency Learning (pT-Learning) framework to estimate an optimal regime that is adaptively adjusted between deterministic and stochastic sparse policy models. It can be further simplified and can easily incorporate off-policy data. We study theoretical properties of sparse policy and establish finite-sample bounds on the excess risk and performance error. The proposed method is implemented by our proximalDTR package and is evaluated through extensive numerical experiments.

Doubly Robust Calibration of Prediction Sets under Covariate Shift

Yachong Yang (University of Pennsylvania), Arun Kumar Kuchibhotla (Carnegie Mellon University), Eric Tchetgen Tchetgen (University of Pennsylvania)*

Conformal prediction has received tremendous attention in recent years and has offered new solutions to problems in missing data and causal inference; yet these advances have not leveraged modern semiparametric efficiency theory for more robust and efficient uncertainty quantification. In this paper, we consider the problem of obtaining distribution-free prediction regions for counterfactuals and individual treatment effects. Under the unconfoundedness assumption, we propose three variants of a general framework to construct well-calibrated prediction regions for the unobserved counterfactuals.

Our approach is based on the efficient influence function for the quantile of the unobserved outcome combined with an arbitrary machine learning prediction algorithm. Next, we extend our approach to account for unobserved confoundedness in a semiparametric sensitivity analysis. We establish that the resulting prediction sets eventually attain nominal average coverage in large samples, which is a consequence of the product bias form of our proposal which implies correct coverage if either the propensity score or the conditional distribution of the response is estimated sufficiently well.

A reduced rank regression model for microbiome-omics data integration

Ying Dai (Oregon State University), Duo Jiang (Oregon State University), Thomas Sharpton (Oregon State University)*

Regression analysis that integrates microbiome data with another omics data type (e.g. metabolomics) is challenging due to the high-dimensionality of both data types. It requires simultaneous estimation of a massive number of association parameters. In regression models where both the response and the explanatory data are high-dimensional, reduced rank regression (RRR) is often useful as it reduces the effective number of parameters by assuming a low-rank coefficient matrix. However, current RRR methods do not capture the compositionality of microbiome data. More specifically, the true abundances of microbes are unobservable, and microbiome composition is only characterized by the relative abundance of a microbe relative to other microbes. To address this challenge, we propose a novel RRR method tailored for identifying the effects on the unobserved microbiome true abundances, while only requiring relative abundance data. We also provide an iterative algorithm guaranteed to attain a global optimum. Simulation studies demonstrate that the proposed method outperforms standard RRR in terms of both estimation precision and prediction accuracy.

Student Paper 7

Exploring Author Roles in Biomedical Publication Networks Using Interactive Visualization

Karen Kanaster (Colorado School of Public Health, University of Colorado Anschutz Medical Campus)

The goal of network-based team science analysis is to identify patterns in topology and connectivity formed through collaboration in scientific research. By integrating personnel data for the biomedical sciences training programs with publication data from PubMed, we have constructed a co-authorship network for the University of Colorado Anschutz Medical Campus for publication years 2010-2020. To flexibly support tasks related to team science analysis, we designed and implemented an interactive network visualization tool with filtering and subnetwork extraction features that allow for exploration and discovery at multiple levels of the collaboration network. The tool provides a network overview along with local and individual-level networks that reveal the different roles and processes involved in network formation. We explored the network through different lenses, including research disciplines, comparisons between different network levels, and egocentric networks. Specifically, we used a combination of social network analysis metrics and visual examination of individual co-authorship networks over time to investigate the collaboration styles of several prominent network members.

Model-Based Voronoi Linkage between Point-Referenced Data and Areal Data in Spatial Analysis with Application to Brazilian Election 2018

Lucas da Cunha Godoy (Department of Statistics, University of Connecticut), Marcos Oliveira Prates (Department of Statistics, Universidade Federal de Minas Gerais), Jun Yan (Department of Statistics, University of Connecticut)*

In Brazil, socioeconomic data are available at census tracts (polygons), while the election data are available at point-referenced voting locations. The misaligned data make it challenging to study the association between electoral and socioeconomic variables. Given that electors are assigned to the nearest electoral sections, we use the Voronoi tessellation to associate each voting station with a polygon. Socioeconomic variables for each polygon are then constructed from such data at the census

tract level assuming that both sets of areal data are driven by the same underlying Gaussian random field. Estimation of the model parameters is done with maximum likelihood. Data for the Voronoi cells are derived from the underlying Gaussian random field with the estimated parameters. A nonparametric alternative approach uses areal interpolation to obtain data for the Voronoi cells from the census tract data. Our simulation study shows that the parametric method is robust in prediction under model misspecification. In application to the election results of Belo Horizonte we have observed that more deprived areas have higher shares of undecided electors.

A Computationally Efficient Approach to Fully Bayesian Benchmarking

Taylor Okonek (University of Washington, Department of Biostatistics), Jon Wakefield (University of Washington, Department of Statistics & Department of Biostatistics)*

In small area estimation, it is often necessary to resort to model-based methods to produce estimates in areas with little or no data. In many settings, we require that some aggregate of small area estimates agree with a national level estimate that may be considered more reliable, for internal consistency. The process of enforcing this agreement is referred to as benchmarking. Few existing benchmarking methods are ideal for applications with non-normal outcomes, and many are computationally inefficient. Fully Bayesian benchmarking is an appealing approach insofar as we can obtain posterior distributions conditional on a benchmarking constraint. However, existing implementations may be computationally prohibitive. We summarize existing benchmarking methods and their shortcomings in a small area estimation setting with binary outcomes, and propose an approach in which an unbenchmarked method that produces samples is combined with a rejection sampler to produce fully Bayesian benchmarked estimates in a computationally efficient way. To illustrate our approach, we provide comparisons of various benchmarking methods in an application to HIV prevalence estimation.

Regression Modeling of Network-Structured Count Data with Exchangeable Dependencies

Wenqin Du (Colorado State University), Wen Zhou (Colorado State University), Bailey K. Fosdick (Colorado State University)*

Statistical analysis on networks has flourished in last decades. While modeling the connectivity among nodes has been broadly studied, the efforts on modeling directed edges with count measurements, and the edgewise dependence only scatter in literature. This paper introduces a novel latent multiplicative Poisson model for directed networks with count edges, where the edgewise dependence of counts is directly modeled by the dependence of latent errors, which is assumed to be weakly exchangeable. The assumption of weak exchangeability covers a variety of commonly-encountered network effects and leads to a concise representation of the error covariance. In addition, identification and inference of the mean structure and regression coefficients depend on the errors only through their covariance. This provides substantial flexibility for our model. We propose a pseudo-likelihood based estimator for the regression coefficients that enjoys asymptotic normality and evaluate our method by extensive numerical studies that corroborate the theory. Our model is then applied to a well-known friendship network data to reveal interesting network effects that are further verified in literature.

Identify shortcomings of Estimators of Discriminative Performance in Time-to-Event Analyses: A Comparison Study

Ying Jin (University of Colorado, Anschutz Medical Campus)*

Modelling time to event outcomes is a major area of methodologic development in biostatistical research. While several estimators have been proposed to assess discriminative performance of such models, including time-dependent AUC and concordance, there exists a previously unidentified feature of a class of estimators which renders them inappropriate for use in many contexts. Specifically, semiparametric estimators have the potential to substantially overestimate out-of-sample discriminative performance. In this paper, we identify the source of this phenomena and illustrate the poor behavior of semiparametric estimators through simulation study and data application. The results

show that out-of-sample estimates of semi-parametric estimators are implausibly higher than in-sample estimates when underlying models overfit to the data, with the degree of divergence increases with increasing model overfit. To address this issue, we additionally propose alternative nonparametric estimators to be used in practice that correctly reflect the true discriminative power of underlying models and recommend smoothing using additive regression models to reduce their high variability.

Student Paper 8

Flexible estimation of the conditional survival function via observable regression models

Charles Wolock (University of Washington), Noah Simon (University of Washington), Marco Carone (University of Washington)*

The conditional survival function of a time-to-event outcome subject to censoring and truncation is a common target of estimation in survival analysis. This parameter may be of scientific interest and also often appears as a nuisance in semiparametric settings. In addition to classical parametric and semiparametric methods (e.g. the proportional hazards model), flexible machine learning approaches have been developed to estimate the conditional survival function. However, many of these methods are targeted toward risk stratification rather than function estimation. Others apply only to discrete time settings or require inverse probability of censoring weights, which can be as difficult to estimate as the outcome survival function itself. Here, we propose novel decompositions of the conditional survival function in terms of observable regression models in which censoring and truncation play no role. This allows application of an array of flexible regression and classification methods rather than only approaches that explicitly handle the complexities inherent to survival data. We outline estimation procedures based on these decompositions and assess their performance via simulation.

Nonparametric Estimation of the Potential Impact Fraction and Population Attributable Fraction

Colleen Chan (Yale University), Rodrigo Zepeda-Tello (National Institute of Public Health of Mexico), Dalia Camcho-García-Formentí (National Institute of Public Health of Mexico), Donna Spiegelman (Yale School of Public Health), Tonatiuh Barrientos-Gutíerrez (National Institute of Public Health of Mexico), Xin Zhou (Yale School of Public Health)*

The estimation of the potential impact fraction (including the population attributable fraction) with continuous exposure data frequently relies on strong distributional assumptions. However, these assumptions are often violated if the underlying exposure distribution is unknown or if the same distribution is assumed across time or space. Nonparametric methods to estimate the potential impact fraction are available for cohort data, but no alternatives exist for cross-sectional data. In this article, we discuss the impact of distributional assumptions in the estimation of the population impact fraction, showing that under an infinite set of possibilities, distributional violations lead to biased estimates. We propose nonparametric methods to estimate the potential impact fraction for aggregated (mean and standard deviation) or individual data (e.g. observations from a cross-sectional population survey), and develop simulation scenarios to compare their performance against standard parametric procedures. We illustrate our methodology on an application of sugar-sweetened beverage consumption on type 2 diabetes incidence. We also present an R package pifpaf to implement these methods.

Long-term effect estimation when combining clinical trial and

Gang Cheng (University of Washington), Yen-Chi Chen (University of Washington), Joseph M. Unger (Fred Hutchinson Cancer Research Center), Cathee Till (Fred Hutchinson Cancer Research Center), Ying-Qi Zhao (Fred Hutchinson Cancer Research Center)*

Combining experimental and observational follow-up datasets has received a lot of attention lately. In a time-to-event setting, recent work has used medicare claims to extend the follow-up period for participants in a prostate cancer clinical trial. This allows the estimation of the long-term effect that cannot be estimated by clinical trial data alone. In this paper, we study the estimation of long-term

effect when participants in a clinical trial are linked to an observational follow-up dataset. Such data linkages are often incomplete for various reasons. We formulate incomplete linkages as a missing data problem with careful considerations of the relationship between the linkage status and the missing data mechanism. We use the popular Cox model as a working model to define the long-term effect. We propose a conditional linking at random (CLAR) assumption and an inverse probability of linkage weighting (IPLW) estimator. We show that our IPLW estimator is consistent and asymptotically normal. We further extend our approach to incorporate time-dependent covariates. Simulation results confirm the validity of our method, and we further apply our methods to the SWOG study.

Evaluating and Improving Methods for Estimating the Effective Reproduction Number

Isaac Goldstein (Department of Statistics, UC Irvine), Jon Wakefield (Departments of Biostatistics and Statistics, University of Washington), Vladimir Minin (Department of Statistics, UC Irvine)*

The effective reproduction number, the average number of individuals a newly infected person will infect, is an important descriptor of an infectious disease epidemic. Counts of newly infected individuals (cases) are a real time indicator of changes in the reproduction number, but estimating the effective reproduction number using cases is challenging because counts can fluctuate due to factors unrelated to underlying disease dynamics, such as testing eligibility and testing supply. We develop a branching process inspired model which incorporates diagnostic test counts as a surveillance model covariate and demonstrate via simulations how incorporating test data allows us to successfully estimate the reproduction number using case data. We show that incorporating tests, as well as other modeling choices we make lead to more precise and accurate estimates as compared to state of the art models. We apply our new model to data from the SARS-CoV-2 epidemic in 15 geographically representative California, USA, counties, and find it produces epidemiologically more plausible estimates of the effective reproduction number as compared to estimates from existing models.

Disease Analytics: COVID-19 Across Canada

Matthew Parker (Simon Fraser University), Jiguo Cao (Simon Fraser Univserity), Laura Cowen (University of Victoria), Lloyd Elliott (Simon Fraser University), Junling Ma (University of Victoria)*

We have developed a new multi-site model for disease analytics. This model uses publicly available disease counts data such as observed cases, recoveries among observed cases, and total deaths. These counts are used to estimate probability of recovery and probability of death among infected individuals, as well as several important population parameters over time including rate of spread, importation of external cases, and case detection probability. We also estimate the total number of active disease cases per region for each reporting interval. We validate the model through simulation studies, which indicate that all model parameters are identifiable. We apply the multi-site model to Canada as a whole, with each province and territory acting as an individual site. Our Canada model estimates the total COVID-19 burden for a large span of the pandemic, 90 weeks from 2020-04-02 to 2022-02-10.

Student Paper 9

A data adaptive rank-based procedure for assessing reproducibility of high-throughput experiments.

Austin Ellingworth (Department of Statistics, Colorado State University), Dr. Debashis Ghosh (Department of Biostatistics & Informatics, Colorado School of Public Health), Dr. Wen Zhou (Department of Statistics, Colorado State University)*

Reproducibility guarantees the consistency and validity of experimental findings. In high-throughput studies reproducibility has often been defined as hypotheses with coinciding test results across experiments. In Philtron et al. (2018), the maximum rank statistic (MaRR) was introduced to identify reproducible hypotheses. Regardless of its empirical success, the theoretical guarantees of MaRR remain largely unknown. We carefully investigate the properties of MaRR which lend it to quantifying reproducibility. Motivated from MaRR, we develop a novel data adaptive statistic that balances a

hypothesis's signal strength and variation across experiment. Based on the new statistic, we design a procedure to identify reproducible hypotheses with marginal false discovery rate control. We show that the new procedure dominates the MaRR statistic with superior power. With a bivariate Gaussian model, we present a revealing phase transition of our procedure. Using simulations, we show the finite sample performance of our method, corroborating our theoretical findings. We also apply our method to two large-scale GWAS of coronary artery diseases in different cohorts and identify reproducible SNPs.

Efficiency loss with binary pre-processing of continuous monitoring data

Paula Langner (University of Colorado Anschutz Medical Campus), Elizabeth Juarez-Colunga (University of Colorado Anschutz Medical Campus), John Rice (University of Colorado Anschutz Medical Campus), Gary Grunwald (University of Colorado Anschutz Medical Campus)*

In studies with a repeatedly measured recurrent event outcome, events may be captured as counts during subsequent intervals or follow-up times either by design or for ease of analysis. In many cases, recurrent events may be further coarsened such that only an indicator of one or more events in an interval is observed at the follow-up time. We examine efficiency loss when coarsening longitudinally observed counts to binary indicators and aspects of the design which impact the ability to estimate a treatment effect of interest. The investigation is motivated by a study of patients with Cardiac implantable electronic devices (CIEDs) in which investigators aimed to examine the effect of treatment on events detected by the devices over time. To study components of such a recurrent event process impacted by data coarsening, we derive the asymptotic relative efficiency (ARE) of a treatment effect estimator utilizing a count outcome, which represents a longitudinal recurrent event process, relative to a coarsened binary outcome. We compare the efficiencies and consider conditions where the binary process maintains good efficiency in estimating a treatment effect.

Efficient Algorithms and Implementation of a Semiparametric Joint Model for Longitudinal and Competing Risk Data: With Applications to Massive Biobank Data

Shanpeng Li (University of California, Los Angeles), Ning Li (University of California, Los Angeles), Hong Wang (Central South University), Jin Zhou (University of California, Los Angeles), Hua Zhou (University of California, Los Angeles), Gang Li (University of California, Los Angeles)*

Semiparametric joint models of longitudinal and competing risk data are computationally costly, and their current implementations do not scale well to massive biobank data. This paper identifies and addresses some key computational barriers in a semiparametric joint model for longitudinal and competing risk survival data. By developing and implementing customized linear scan algorithms, we reduce the computational complexities from O(n^2) or O(n^3) to O(n) in various steps including numerical integration, risk set calculation, and standard error estimation, where is the number of subjects. Using both simulated and real-world biobank data, we demonstrate that these linear scan algorithms can speed up the existing methods by a factor of up to hundreds of thousands when , often reducing the runtime from days to minutes. We have developed an R package, FastJM, based on the proposed algorithms for joint modeling of longitudinal and competing risk time-to-event data and made it publicly available on the Comprehensive R Archive Network (CRAN).

Fast Distributed Principal Component Analysis of Large-Scale Federated Data

Shuting Shen (Harvard University), Junwei Lu (Harvard University), Xihong Lin (Harvard University)*

Principal component analysis (PCA) is one of the most popular methods for dimension reduction. In light of rapidly increasing large-scale data in federated ecosystems, the traditional PCA method is often not applicable due to privacy protection consideration and large computational burden. Fast PCA algorithms have been proposed to lower the computational cost but cannot handle federated data. Distributed PCA algorithms have been developed to handle federated data but are not computationally efficient when data at each site are very large. In this paper, we propose the FAst DIstributed (FADI) PCA method which applies fast PCA to site-specific data using multiple random sketches and aggregates the results across

sites. We perform a non-asymptotic theoretical study to show that FADI enjoys the same error rate as the traditional full sample PCA and a much smaller order of computational burden compared to existing methods. We perform extensive simulation studies and show that FADI substantially outperforms the other methods in computational efficiency without sacrificing statistical accuracy. We apply FADI to the analysis of the 1000 Genomes data to study the population structure.

Regression in Tensor Product Spaces by the Method of Sieves

Tianyu Zhang (University of Washington), Noah Simon (University of Washington)*

Many statistical problems require estimating an unknown underlying function (e.g., the conditional mean function) that best links a set of features to an outcome of interest. On the one hand, many applications favor more flexible (nonparametric) procedures for consistent estimates; on the other hand, effective estimation in some popular (isotropic) function spaces can be prohibitively difficult -- both statistically and computationally -- especially when the number of features is large. In this paper, we present some sieve estimators for regression in nonparametric tensor product spaces. This type of models ``scale" better with the feature dimension than many classical spaces. At the same time, the corresponding sieve estimators can be easily applied to multivariate nonparametric problems and have appealing statistical and computational properties. Moreover, they can effectively leverage additional structures such as feature sparsity. In addition to the theoretical guarantees, we also present numerical examples to compare the finite-sample performance of the proposed estimators with several popular machine learning methods.