Michigan Student Symposium for Interdisciplinary Statistical Sciences

MSSISS 2021

Feb 25th, 2021

1:00 pm – 6:30 pm

Feb 26th, 2021

8:55 am – 4:30 pm

Virtual Meeting

Sponsored by: American Statistical Association (ASA), Michigan Institute for Data Science (MIDAS), UM Survey Research Center (SRC), UM Rackham Graduate School, Department of Biostatistics, Statistics, Electrical Engineering & Computer Science, Industrial & Operations Engineering, and Program in Survey Methodology

Table of Contents

Committee and Acknowledgements	2
Schedule	3
Keynote Speaker, Friday: Dr. Xihong Lin	5
Michigan Junior Faculty Speaker, Thursday: Dr. Raed Al Kontar	6
Abstracts, Thursday	
Speed Oral Presentations	7
Poster Session I	12
Oral Presentations I	17
Abstracts, Friday	
Oral Presentations II	20
Poster Session II	23
Oral Presentations III	28

Committee & Acknowledgments

Sponsoring Departments: Biostatistics, Electrical Engineering & Computer Science, Industrial & Operations Engineering, Statistics, Survey Methodology

MSSISS 2021 Student Organizing Committee:

Yijun Li (Department of Biostatistics)
Zeyu Sun (Department of Electrical Engineering & Computer Science)
Seokhyun Chung (Department of Industrial & Operations Engineering)
Ziping Xu (Department of Statistics)
Xinyu Zhang (Program in Survey Methodology)

MSSISS 2021 Faculty Advisory Committee:

Dr. Kang Jian Johnson (Department of Biostatistics)
Dr. S. Sandeep Pradhan (Department of Electrical Engineering & Computer Science)
Dr. Raed Al Kontar (Department of Industrial & Operations Engineering)
Dr. Edward Ionides (Department of Statistics)
Dr. Brady West (Program in Survey Methodology)

We offer our sincere thanks to each member of the faculty committee for their useful suggestions in planning the MSSISS conference, as well as to last year's committee for their insight.

We are grateful to the Michigan Institute for Data Science (MIDAS), the Rackham Graduate School, and Survey Research Center (SRC) for their generous support in sponsoring our event.

We are grateful to the Ann Arbor Chapter of the American Statistical Association for their generous support in providing the ASA-sponsored prize for best poster of interdisciplinary application.

MSSISS 2021

Official Schedule: Thursday, February 25th

Recurrent Zoom Meeting ID: 979 7179 8346 / Poster Session on Gather town

1:00 – 1:05pm Welcoming Remarks (Zoom) (Moderator: Yijun Li)

1:05 – 2:00pm Speed Oral Presentations (Zoom) (Moderator: Ziping Xu)

Emily Roberts, Department of Biostatistics (1:05 - 1:10pm)
Validating Surrogate Endpoints with Longitudinal Outcomes

Charlotte Z Mann, Department of Statistics (1:10 - 1:15pm)
Derivation and external validation of a simple risk score to predict in-hospital mortality in patients hospitalized for COVID-19

Debarghya Mukherjee, Department of Statistics (1:15 - 1:20pm)
On the efficient estimation of Regression Discontinuity Design in presence of confounding factors

Sijia Geng, Department of EECS (1:25 - 1:30pm)
A Data-Driven Approach for Self-Tuning of Power System Stabilizer Using Graph Neural Network

Yijun Li, Department of Biostatistics (1:30 - 1:35pm)
Evaluating Data Integration Tools on Clustering Spatial Transcriptomics Data

Fatema Shafie Khorassani, Department of Biostatistics (1:35 - 1:40pm)
Incorporating Patient Subgroups in Surrogate Paradox Measures

Advaidh Venkat, Department of IOE (1:40 - 1:45pm)
Evaluating Patient Triage Strategies for Non-Emergency Outpatient Procedures Under Reduced Capacity Due to the COVID-19 Pandemic

Alexander Ritchie, Department of EECS (1:45 - 1:50pm)
Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations

2:00 – 3:00pm Poster Session I + Speed Presentation Poster Session (Gather town)

3:00 – 4:50pm Oral Presentations I (Zoom) (Moderator: Zeyu Sun)

Yiwang Zhou, Department of Biostatistics (3:00 - 3:20pm)
Synergistic Self-Learning Approach to Establishing Personal Nutrition Intervention Schemes from Multiple Benefit Outcomes in a Calcium Supplementation Trial

Xubo Yue, Department of IOE (3:20 - 3:40pm)
R'enyi Variational Inference for Gaussian Processes: Towards Enabling Regularization Tuning

Wenbo Wu, Department of Biostatistics (3:40 - 4:00pm)
Analysis of Readmissions Data Taking Account of Competing Risks

Daiwei Zhang, Department of Biostatistics (4:00 - 4:20pm)
Image-on-Scalar Regression via Neural Networks

Yujia Pan, Department of Statistics (4:20 - 4:40pm)
Separating and integrating latent variables for improved classification of genomic data

4:50 – 5:55pm Michigan Junior Faculty Keynote (Zoom) (Moderator: Seokhyun Chung)

Assistant Professor Raed AI Kontar, Department of IOE, University of Michigan
Predictive Analytics for IoT Enabled Systems

5:55 – 6:30pm Closing Remarks (Zoom) (Moderator: Seokhyun Chung)

Official Schedule: Friday, February 26th

Recurrent Zoom Meeting ID: 979 7179 8346 / Poster Session on Gather town

8:55 – 9:00am Welcoming Remarks (Zoom) (Moderator: Xinyu Zhang)

9:00 – 10:40am Oral Presentations II (Zoom) (Moderator: Xinyu Zhang)

Peter MacDonald, Department of Statistics (9:00 - 9:20am)

Latent space models for multiplex networks with shared structure

Robert P Malinas, Department of EECS (9:20 - 9:40am)

Space-Time Adaptive Detection at Low Sample Support

Emily Morris, Department of Biostatistics (9:40 - 10:00am)

Scalar on Network Regression via Boosting

Michael Law, Department of Statistics (10:00 - 10:20am)

Functional Data Analysis with Ultra High-Dimensional Covariates

Tian Gu, Department of Biostatistics (10:20 - 10:40am)

A Bayesian stacked imputation framework for extended regression inference in data integration

10:50 – 11:50am Poster Session II (Gather town)

12:00 – 1:30pm Lunch Break

1:30 – 3:10pm Oral Presentations III (Zoom) (Moderator: Yijun Li)

Xinyu Zhang, Program in Survey Methodology (1:30 - 1:50pm)

A Multivariate Stopping Rule for Survey Data Collection

Guangyu Yang, Department of Biostatistics (1:50 - 2:10pm)

Estimation of Knots in Linear Spline Model

Zihong Yi, Department of IOE (2:10 - 2:30pm)

A Multi-batch L-BFGS Method with Variance Reduction: Theory and Experiments

Daniel Kessler, Department of Statistics (2:30 - 2:50pm)

Inference post Selection of Group-sparse Regression Models

Drew Yarger, Department of Statistics (2:50 - 3:10pm)

ArgoSSM: A Bayesian state-space framework for predicting the location of oceanographic sensors in the Southern Ocean

3:10 – 4:00pm Keynote Address (Zoom) (Moderator: Yijun Li)

Professor Xihong Lin, Harvard University

Learning from COVID-19 Data on Transmission, Health Outcomes, Interventions and Vaccination

4:00 – 4:30pm Student Awards and Closing Remarks (Zoom) (Moderator: Zeyu Sun)

Keynote Speaker, Friday: Dr. Xihong Lin

Dr. Xihong Lin is a highly accomplished Statistician and Scientist. Dr. Lin is Professor and former Chair of the Department of Biostatistics, Coordinating Director of the Program in Quantitative Genomics at the Harvard T. H. Chan School of Public Health, Professor of the Department of Statistics at the Faculty of Arts and Sciences of Harvard University, and Associate Member of the Broad Institute of Harvard and MIT. Dr. Lin holds many prestigious titles and awards. Dr. Lin is an elected member of the National Academy of Medicine. She received the 2002 Mortimer Spiegelman Award from the American Public Health Association, the 2006 Committee of Presidents of Statistical Societies (COPSS) Presidents' Award, and the 2017 COPSS FN David Award. Dr. Lin is also an elected fellow of the American Statistical Association (ASA), Institute of Mathematical Statistics, and International Statistical Institute. Dr. Lin's research focuses on developing and applying statistical and computational methods to analyze massive data from the genome, exposome, and phenome. She is also interested in scalable statistical inference and learning for big genomic, epidemiological, and health data. Dr. Lin's statistical methodological research has been supported by the MERIT Award (R37) (2007-2015) and the Outstanding Investigator Award (OIA) (R35) (2015-2022) from the National Cancer Institute (NCI).

Title: Learning from COVID-19 Data on Transmission, Health Outcomes, Interventions and Vaccination

COVID-19 is an emerging respiratory infectious disease that has become a pandemic. In this talk, I will first provide a historical overview of the epidemic in Wuhan. I will then provide the analysis results of 32,000 lab-confirmed COVID-19 cases in Wuhan to estimate the transmission rates using Poisson Partial Differential Equation based transmission dynamic models. This model is also used to evaluate the effects of different public health interventions on controlling the COVID-19 outbreak, such as social distancing, isolation and quarantine. I will present the results on the epidemiological characteristics of the cases. The results show that multi-faceted intervention measures successfully controlled the outbreak in Wuhan. I will next present transmission regression models for estimating transmission rates in USA and other countries, as well as factors including intervention effects using social distancing, test-trace-isolate strategies that affect transmission rates. I will discuss estimation of the proportion of undetected cases, including asymptomatic, pre-symptomatic cases and mildly symptomatic cases, the chances of resurgence in different scenarios, and the factors that affect transmissions. I will also present the US county-level analysis to study the demographic, social-economic, and comorbidity factors that are associated with COVID-19 case and death rates. I will also present the analysis results of >500,000 participants of the HowWeFeel project on symptoms and health conditions in US, and discuss the factors associated with infection, behavior, and vaccine hesitancy. I will provide several takeaways and discuss priorities.

Michigan Junior Faculty Speaker, Thursday: Dr. Raed Al Kontar

Dr. Raed Al Kontar is an Assistant Professor in the Industrial & Operations Engineering department at the University of Michigan, Ann Arbor. He received his Ph.D. in Industrial and Systems Engineering in 2018 and M.S in Statistics in 2017 from the University of Wisconsin Madison. He also received his B.S in Civil and Environmental Engineering with a minor in Mathematics from the American University of Beirut (AUB) in 2014. Raed's main research interest is data science using probabilistic models where he aims to understand the foundations of such models in extracting interpretable knowledge and generalizing to new data. Raed also focuses on data science applications within Internet of Things (IoT) enabled systems, specifically in tele-service settings. Raed has recently received best paper awards at INFORMS quality statistics & reliability section (2018, 2019) and data mining section (2020), and the Quality Control and Reliability Engineering section (2019) from IISE.

Title: Predictive Analytics for IoT Enabled Systems

Internet of things (IoT) enabled systems have become increasingly available in practice. Examples include GM's OnStar® tele-service system, the InSite® telemonitoring system from GE, smart home appliances, and various personalized remote patient monitoring systems. The unprecedented data availability in such connected systems has ushered in the present-day era of Industry 4.0, where smart data analytics drive smart decisions. In this talk, I focus on predictive analytics and discuss both opportunities and challenges that IoT presents in building a unified predictive framework. I then shed light on the state-of-the-art efforts to improve generalization performance of both kernel-based and deep learning predictive models.

Speed Oral Presentations

3A. Validating Surrogate Endpoints with Longitudinal Outcomes
Emily Roberts, PhD Candidate, Department of Biostatistics
Co-Authors: Michael Elliott, Jeremy Taylor
Keywords: cross-over design, longitudinal outcomes, surrogate endpoints, principal stratification, Bayesian methods

In the context of a clinical trial, several methods have been proposed to validate surrogate endpoints in place of the true outcome of interest. We extend the previously proposed principal surrogacy framework based on the causal effect predictiveness surface and subsequent work by Conlon et al. (Biostatistics, 2014) to model the joint distribution of normally-distributed potential outcomes. In this work, we incorporate longitudinal measurements of the true outcomes using a mixed modeling approach. We compare the estimation properties of a fully Bayesian imputation method to that of using the observed data only and explore the impact of different prior distributions on non-identified parameters. We further consider possible conditional independence of the random effects. Finally, we consider a surrogate-dependent treatment efficacy curve and how we can validate the surrogate at different time points or integrate over several for an overall conclusion of surrogacy. Our work is motivated by a clinical trial of a gene therapy where the functional outcomes are measured repeatedly throughout the trial. This trial also utilizes a cross-over design where all patients eventually receive the treatment. We update our models to incorporate this additional information.

3B. Derivation and external validation of a simple risk score to predict in-hospital mortality in patients hospitalized for COVID-19
Charlotte Z. Mann, PhD Pre-candidate, Department of Statistics
Co-Authors: Ben Hansen, Lauren Gaydosh
Keywords: COVID-19, risk score model, AUC, model selection

As SARS-CoV-2 continues to spread, and hospitals are treating a large number of patients with COVID-19, easy-to-use risk models that predict hospital mortality can assist in clinical decision making and triage. We aimed to develop a risk score model for in-hospital mortality in patients hospitalized for COVID-19 that was robust across hospitals and used clinical factors that are readily available and measured standardly across hospitals. Employing an iterative forward selection approach that involved quantitative assessment as well as physician expertise, we developed a risk score model using observational data collected by professional abstractors for patients in 20 diverse hospitals across the state of Michigan. External validation of the model on 19 Michigan hospitals not included in the derivation showed high discrimination with an AUC of .8. We conclude that risk of in-hospital mortality in COVID-19 patients can be estimated with high discrimination using a few factors, which are standardly measured and available to physicians very early in a hospital encounter.

3C. On the efficient estimation of Regression Discontinuity Design in presence of confounding factors

Debarghya Mukherjee, PhD Candidate, Department of Statistics
Co-Authors: Moulinath Banerjee, Ya'acov Ritov
Keywords: Regression discontinuity design, treatment effect, semiparametric efficiency

Regression discontinuity design is a widely used method to assess treatment effects in psychology, psychometry and statistics. This method is primarily employed when a treatment is assigned to an individual based on some of their characteristics (e.g. scholarship is allocated based on merit) instead of random allocation. Popular methods that have been largely employed till date to estimate the treatment effect suffer from a slow rate of convergence (i.e. slower than the gold standard root-n rate, n being the number of sample) owing to certain localizations involved in such procedures. In this paper, we present a novel methodology which enables us to estimate the treatment effect at root-n rate in the presence of fairly general forms of confoundedness. Moreover, we show that our estimator is semi-parametrically efficient . We analyze two real datasets using our method: the first is to understand the effect of Islamic ruling party on the women empowerment in Turkey (based on the data from 1994 municipal election) and the other is to analyze the effect of probation on the subsequent grades of the students (based on the data from anonymous Canadian university).

3D. A Data-Driven Approach for Self-Tuning of Power System Stabilizer Using Graph Neural Network

Sijia Geng, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Jiayang Xu
Keywords: Graph Neural Network, Power System Stabilizer, Distributed Energy Resources, Data-Driven Approach

As the fast depletion of fossil fuels and the increasing concern on environmental issues, more and more renewable distributed energy resources (DERs) have been integrated into power systems. DERs are usually integrated into power systems through power electronics, for example, the so-called voltage source converter (VSC), which demonstrate drastically different dynamical behaviors than the traditional synchronous generator (SG). SG provides damping, i.e., a favorable stabilizing feature, to the power system, which VSC does not possess. Moreover, for a traditional power system which is dominated by SGs, the system operation point does not alter significantly. However, renewable resources such as wind and solar exhibit large variability over the day, which will inevitably change the operating points of traditional SGs and other devices in the power system. These characteristics accompanying the VSC-interfaced DERs are starting to pose significant challenges to power system operators.
SGs provide damping to the power system through a controller, referred to as the power system stabilizer (PSS). Careful design and tuning of PSS are important in order to maintain stable and safe operation of power systems. Traditionally, such tunings are carried out iteratively for all SGs in the system, under a given operating point. In this work, we propose a novel PSS design approach based on graph neural network (GNN), while specifically taking into account the large variation of operating condition caused by renewable DER generation.
There has been extensive work in the literature on designing adaptive self-tuning power system stabilizer. More recent developments explore artificial intelligence techniques such as

convolutional neural network. In this work, we propose a novel design approach by exploring graph neural network (GNN), thereby retain the non-Euclidean data structure and the topology of physical power system. A heterogeneous graph is constructed to represent the power system, where PSS, SG and VSC are represented by parametric nodes. The system dynamics are predicted using the GNN, and a reward function is formulated to measure the stability of the predicted response. Reinforcement learning (RL) is used to train an agent to maximum the reward by tuning the PSS control parameters.
It is expected that rapid growth of deployment of data measuring units, such as phase measurement units, will be available in power systems. We exploit the data measurements such as real and reactive powers on the power system nodes to train and validate the GNN model. The proposed designing framework is evaluated on a power system model, under fast varying DER generation profiles, and statistical test results are established.

3E. Evaluating Data Integration Tools on Clustering Spatial Transcriptomics Data

Yijun Li, PhD Candidate, Department of Biostatistics
Co-Authors: Bing He, Zheng Jing, Qianhui Huang
Keywords: Spatial Transcriptomics, feature extraction, clustering, data integration, benchmarking

In recent years, researchers have developed various technologies allowing for sequencing spatially resolved transcriptome data. The spatial patterns of expression variation, represented by spatially variable genes, are unique features in spatial transcriptome data. However, these spatially variable genes are often very different from the same dataset, Äôs highly variable genes, which are selected by their expression values and typically used to cluster cell types. Therefore, the integration of spatially variable genes and highly variable genes may improve cell type clustering accuracy solely based on highly variable genes. Towards this, we compared five different computational methods including LIGER, Seurat v3, Principal Component Analysis (PCA), and MOFA+ to integrate spatially variable and highly variable genes, followed by extracting reduced features for clustering. We applied these methods to two real datasets from two different spatial transcriptomics technologies: Mouse Somatosensory Cortex data collected via SeqFISH+ and Mouse Olfactory Bulb data collected using St√•hl et al.,Äôs Spatial Transcriptomics. We applied different clustering methods to the results of these models and evaluated their performance in both non-spatial and spatial context, using priorly identified cell types. Additionally, we also conducted simulations to evaluate the robustness of these methods under different patterns of transcriptomics and spatial expression. To our knowledge, this is the first benchmarking effort to evaluate the gain of accuracy on combining different feature sets in spatial transcriptomics data, through different data integration tools.

3F. Incorporating Patient Subgroups in Surrogate Paradox Measures

Fatema Shafie Khorassani, PhD Candidate, Department of Biostatistics
Co-authors: Jeremy M.G. Taylor, Nico Kaciroti, Michael R. Elliott
Keywords: Surrogate endpoint, surrogate paradox, clinical trials, subgroups

Clinical trials often collect intermediate or surrogate endpoints other than their true endpoint of interest. It is important that the treatment effect on the surrogate endpoint accurately predicts the treatment effect on the true endpoint. There are settings in which the proposed surrogate endpoint is positively correlated with the true endpoint but the treatment has opposite effects on the surrogate and true endpoints, a phenomenon labeled ‚Äúsurrogate paradox‚Äù. Covariate information may be useful in predicting an individual‚Äôs risk of surrogate paradox. In this work, we propose methods for incorporating covariates into measures of assessing the risk of surrogate paradox using the meta-analytic causal association framework. The measures calculate the probability that a treatment will have opposite effects on the surrogate and true endpoints and determine the size of a positive treatment effect on the surrogate endpoint that would reduce the risk of a negative treatment effect on the true endpoint as a function of covariates. We allow the effects of covariates on the surrogate and true endpoint to vary across trials and describe methods to test that assumption.

3G. Evaluating Patient Triage Strategies for Non-Emergency Outpatient Procedures Under Reduced Capacity Due to the COVID-19 Pandemic

Advaidh Venkat, Undergraduate student, Department of Industrial & Operations Engineering
Co-Authors: Adam VanDeusen, Che-Yi Liao, Amy M. Cohn, Jacob Kurlander, Sameer Saini
Keywords: Simulation Modeling, Discrete-Event Simulation, COVID-19, Healthcare Applications

The COVID-19 pandemic impacted the healthcare system in several ways, including the cancellation or deferral of non-urgent medical appointments due to systems reducing capacity to keep patients safe and abide by governmental orders. We develop a discrete-event simulation to model how a clinical facility under such circumstances may safely triage patients to alternative/delayed appointment options and incrementally add back capacity as restrictions loosen. Our model is applied to colonoscopy procedures at a Veterans Affairs clinic in Ann Arbor, Michigan. We consider patients of different risk categories arriving each week‚Äîhighest risk patients are seen first, and lower risk patients wait in a queue. We develop scenarios of integrating three triage strategies, as described by clinical collaborators, and evaluate metrics such as average patient wait time and number of patients reaching a designated wait time. We find that implementing more triage strategies allows more patients to be seen, while reducing average wait time. Our work serves as a clinical decision-making tool for healthcare facilities facing pandemic-induced capacity reduction and is currently being expanded to model other clinics.

3H. Consistent Estimation of Identifiable Nonparametric Mixture Models from Grouped Observations

Alexander Ritchie, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Robert Vandermeulen, Clayton Scott
Keywords: Nonparametric Density Estimation, Mixture Models

Recent research has established sufficient conditions for finite mixture models to be identifiable from grouped observations. These conditions allow the mixture components to be nonparametric and have substantial (or even total) overlap. This work proposes an algorithm that consistently estimates any identifiable mixture model from grouped observations. Our analysis leverages an oracle inequality for weighted kernel density estimators of the distribution on groups, together with a general result showing that consistent estimation of the distribution on groups implies consistent estimation of mixture components. A practical implementation is provided for paired observations, and the approach is shown to outperform existing methods, especially when mixture components overlap significantly.

Poster Session I

P1a. Bayesian Inferences in EEG-Based Brain-Computer Interface via the Split-and-Merge Gaussian Process

Tianwen Ma, PhD Candidate, Department of Biostatistics
Co-Authors: Yang Li, Jane E. Huggins, Ji Zhu
Keywords: Classification; Bayesian Methods; Gaussian Processes; Brain-computer Interface

A brain-computer interface (BCI) uses brain activity to control technology. BCIs are intended to help people with disabilities use technology for communication. The fundamental statistical problem in BCI is classification. A common design for an electroencephalogram (EEG) BCI relies on classification of the P300 event-related potential (ERP), which is a response elicited by the rare occurrence of target stimuli among common non-target stimuli. Many machine learning methods have constructed P300 ERP-based classifiers, but few try providing insights on the underlying mechanism of the neural activity. In this work, we propose a new Bayesian generative method to model the conditional distribution of the EEG signals given our EEG-BCI design, from which the predictive probability of brain signals can be derived. Our method focuses on detecting spatial-temporal differences where the EEG signals have strong predictive powers, providing an understanding of the neural activity in response to external stimuli. Extensive simulation studies and analysis of real participants show the advantages of the proposed method compared to alternative methods.

P1b. Quiet Students in Teams: Identifying Their Traits and Predicting Them

Jeong Hin Chin, Undergraduate Student, Department of Statistics
Keywords: Quiet students; Bootstrap; Problems in collaborative learning; Predictions; Cross-validation

Collaborative learning(CL) is shown to improve students,Äô academic achievement and help in increasing self-efficacy and experience. Nonetheless, communication difficulties among teammates is the reason why some students think badly of CL. In this paper, we investigate the relationship between teamwork experiences and silence in teams so that instructors are able to sperate students into their best fit teams based on students,Äô beginning of term survey. Specifically, we analyze how holding back ideas or feeling not belonged to a course relates to the extraversion score of a student. By performing Bootstrap on the data collected, the result suggests that students who hold back ideas or feeling not belonged to a course are likely to be quiet in team discussions. However, students with many past teamwork experiences will be more talkative and more likely to fit into a new course that contains teamwork projects. Using cross validation, we find that student,Äôs belongingness to a course is a more accurate indicator in determining quiet students. Lastly, hierarchical clustering is used to cluster the data and help determining the probability of new students in each cluster being quiet.

P1c. Single-cell trajectories can be reconstructed through the calculation of pseudotime from trajectory inference(TI) methods.

Jingyi Zhai, PhD Candidate, Department of Biostatistics
Co-Authors: Hui Jiang
Keywords: Single-cell RNA-sequencing, Cell trajectory reconstruction, Stochastic optimization

The single-cell RNA-sequencing (scRNA-seq) technology is a recent advancement that enables the measurement of gene expression at single cell level, so it prompts the understanding in the dynamic cellular process. Reconstructing a cell trajectory from the gene expression for a sample of cells is introduced as a new research area by this technology. The high-dimensional gene expression data space and the associated high-level noise pose difficulties in modeling the trajectory from the original expression data. We develop a new trajectory reconstruction (TR) method to estimate a tree-structured cell trajectory from scRNA-seq data. We derive a penalized likelihood framework and a stochastic optimization algorithm to search through the non-convex tree space to obtain the global solution. We compare our proposed approach with other existing methods using simulated and real scRNA-seq data sets. As the simulation study and the real data example show, our algorithm is more accurate and less sensitive to outliers than other compared methods in terms of cell ordering estimation.

P1d. Minimax optimal approaches to the label shift problem

Subha Maity, PhD Candidate, Department of Statistics
Co-Authors: Yuekai Sun, Moulinath Banerjee

We study the minimax rates of the label shift problem in non-parametric classification. In addition to the unsupervised setting in which the learner only has access to unlabeled examples from the target domain, we also consider the setting in which a small number of labeled examples from the target domain is available to the learner. Our study reveals a difference in the difficulty of the label shift problem in the two settings, which we attribute this difference to the availability of data from the target domain to estimate the class conditional distributions in the latter setting. We also show that a class proportion estimation approach is minimax rate-optimal in the unsupervised setting.

P1e. Low-Rank Generalized Linear Bandit Problems

Yangyi Lu, PhD Candidate, Department of Statistics
Co-Authors: Ambuj Tewari, Amirhossein Meisami
Keywords: Bandit; Low-Rank

In a low-rank linear bandit problem, the reward of an action (represented by a matrix of size d_1 times d_2) is the inner product between the action and an unknown low-rank matrix Θ^* . We propose an algorithm based on a novel combination of online-to-confidence-set conversion~citep{abbasi2012online} and the exponentially weighted average forecaster constructed by a covering of low-rank matrices. In T rounds, our algorithm achieves $\tilde{O}((d_1 + d_2)^{3/2}\sqrt{rT})$ regret that improves upon the standard linear bandit regret bound of $\tilde{O}(d_1 d_2 \sqrt{T})$ when the rank of Θ^* : $r \, ll \, min\{d_1, d_2\}$. We also extend our algorithmic approach to the generalized linear setting to get an algorithm which enjoys a similar bound under regularity conditions on the link function. To get around the computational intractability of covering based approaches, we propose an efficient algorithm by extending the Explore-Subspace-Then-Refine algorithm of cite{jun2019bilinear}. Our efficient algorithm achieves $\tilde{O}((d_1 + d_2)^{3/2}\sqrt{rT})$ regret under a mild condition on the action set X and the r -th singular value of Θ^* . Our upper bounds match the conjectured lower bound of cite{jun2019bilinear} for a subclass of low-rank linear bandit problems. Further, we show that existing lower bounds for the sparse linear bandit problem strongly suggest that our regret bounds are unimprovable. To complement our theoretical contributions, we also conduct experiments to demonstrate that our algorithm can greatly outperform the performance of the standard linear bandit approach when Θ^* is low-rank.

P1f. Individualized Risk Assessment of Preoperative Opioid Use by Interpretable Neural Network Regression

Yuming Sun, PhD Candidate, Department of Biostatistics
Co-Authors: Jian Kang, Chad Brummett, Yi Li
Keywords: Deep neural network; Interpretable neural network regression; Preoperative opioid use

As preoperative opioid use is associated with worse postoperative outcomes and increased postoperative healthcare utilization and expenditures, understanding the risk of preoperative opioid use helps establish effective opioid management for each patient. In the field of machine learning, deep neural network (DNN) has emerged as a powerful means for risk assessment because of its superb prediction power; however, the blackbox algorithms make the results less interpretable than statistical models. Bridging the gap between the statistical and machine learning fields, we propose a novel Interpretable Neural Network Regression (INNER), which combines the strengths of statistical and DNN models. We use the proposed INNER to conduct individualized risk assessment of preoperative opioid use. Intensive simulations and statistical analysis of 34,186 patients expecting surgery in the Analgesic Outcomes Study (AOS) show that, the proposed INNER accurately predicts the preoperative opioid use based on preoperative characteristics. It also estimates the patient-specific odds of opioid use without pain and the odds ratio of opioid use for one unit increase in the reported overall body pain, leading to more straightforward interpretations on opioid tendency compared to DNN. Our analysis identifies patient characteristics associated with the opioid tendency and is largely consistent with the previous findings, evidencing that INNER is a useful tool for individualized risk assessment of preoperative opioid use.

P1g. Empirical Maximum Likelihood Normalization for Drug Screening Data
Zoe Rehnberg, PhD Candidate, Department of Statistics
Co-Authors: Johann Gagnon-Bartsch

Large-scale pharmacognomic experiments are typically used to study the efficacy and promise of potential anti-cancer drugs and to identify genetic predictors of drug sensitivity. In these studies, data on drug efficacy, the drug screening data, is collected on tens of thousands of microplanes throughout the duration of the experiment, which can last several years. To effectively use the raw drug screening data and to compare the performance of drugs across plates and across studies, raw measurements must be normalized. This process produces relative viabilities that should ideally be comparable, regardless of experimental site, date, or plate. While most analyses use control wells to do the normalization, typical methods fail to take into account technical errors, including complex spatial biases, that interfere with the accuracy of normalization. In this work, we develop a new method of normalizing raw drug screening data that accounts for the presence of technical errors. This method takes advantage of both control wells and drugged wells to do the normalization and improves agreement between replicated measurements.

P1h. HePPCAT: Probabilistic PCA for Data with Heteroscedastic Noise
Kyle Gilman, PhD Candidate, Department of Electrical Engineering and Computer Science
Co-Authors: David Hong, Laura Balzano, and Jeffrey Fessler
Keywords: Heteroscedastic PCA, nonconvex optimization, signal processing

Principal component analysis (PCA) is a classical and ubiquitous method for reducing data dimensionality, but it is suboptimal for heterogeneous data that are increasingly common in modern applications. PCA treats all samples uniformly so degrades when the noise is heteroscedastic across samples, as occurs, e.g., when samples come from sources of heterogeneous quality. This paper develops a probabilistic PCA variant that estimates and accounts for this heterogeneity by incorporating it in the statistical model. Unlike in the homoscedastic setting, the resulting nonconvex optimization problem is not seemingly solved by singular value decomposition. This paper develops a heteroscedastic probabilistic PCA technique (HePPCAT) that uses efficient alternating maximization algorithms to jointly estimate both the underlying factors and the unknown noise variances. Simulation experiments illustrate the comparative speed of the algorithms, the benefit of accounting for heteroscedasticity, and the seemingly favorable optimization landscape of this problem.

P1i. Decision Making Problems with Funnel Structure: A Multi-Task Learning Approach with Application to Email Marketing Campaigns

Ziping Xu, PhD Candidate, Department of Statistics
Co-Authors: Amir Meisami, Ambuj Tewari
Keywords: Decision making; Multi-task learning; Email marketing

This paper studies the decision making problem with Funnel Structure. Funnel structure, a well-known concept in the marketing field, occurs in those systems where the decision-maker interacts with the environment in a layered manner receiving far fewer observations from deep layers than shallow ones. For example, in the email marketing campaign application, the layers correspond to Open, Click and Purchase events. Conversions from Click to Purchase happen very infrequently because a purchase cannot be made unless the link in an email is clicked on. We formulate this challenging decision making problem as a contextual bandit with funnel structure and develop a multi-task learning algorithm that mitigates the lack of sufficient observations from deeper layers. We analyze both the prediction error and the regret of our algorithms. We verify our theory on prediction errors through a simple simulation. Experiments on both a simulated environment and an environment based on real-world data from a major email marketing company show that our algorithms offer significant improvement over previous methods.

P1j. Gradient coding with the Hadamard transform

Neophytos Charalambides, PhD Candidate, Department of Electrical Engineering and Computer Science
Co-Authors: Alfred Hero, Mert Pilanci
Keywords: Sampling, gradient descent, coded computing

A major impediment in machine learning is scalability of algorithms to massive datasets. Two ways of overcoming this are to 1) distribute the computations among several workers, and 2) resort to approximate randomized algorithms. A common issue which arises in distributed computing is the presence of stragglers. This issue has been addressed by coding-theory techniques, for both exact and approximate computations. In this paper, we show how one can utilize a uniform sampling idea from randomized numerical linear algebra to obtain an approximate gradient coding scheme for linear regression. Specifically, we utilize the Subsampled Randomized Hadamard Transform.

Oral Presentations I

O1a. Synergistic Self-Learning Approach to Establishing Personal Nutrition Intervention Schemes from Multiple Benefit Outcomes in a Calcium Supplementation Trial

Yiwang Zhou, PhD Candidate, Department of Biostatistics
Co-Authors: Peter X.K. Song
Keywords: Dietary supplement, DOHaD hypothesis, O-learning, Precision nutrition, Support vector machine.

Precision nutrition is an emerging research field in nutritional sciences. Being a major risk to children's neurobehavioral and cognitive development, excessive in utero exposure to lead for embryos would be detrimental if no intervention is in place. The calcium supplementation trial conducted by the ELEMENT team aims to study the effect of daily calcium supplement in reducing maternal lead exposure to infants during pregnancy. This article focuses on establishing a personal nutrition intervention scheme (PNIS) that can guide pregnant women on taking daily calcium supplementation to maximize the reduction of maternal lead exposure to infants. In the analysis, we present a novel method, termed Synergistic Self-learning (SS-learning), to address two major challenges in the derivation of PNIS in the presence of multiple clinical outcomes, including heterogeneous multidimensional outcomes and complex missing data patterns. SS-learning can effectively synergize heterogeneous features of multiple training data sources in the derivation of PNIS. Our analysis of the ELEMENT calcium supplementation trial identified several important predictors used to form a PNIS that would give a higher expected lead reduction should it be implemented to the whole study population. We examined the sensitivity and stability of SS-learning used in the analysis by comprehensive simulation studies.

O1b. R'enyi Variational Inference for Gaussian Processes: Towards Enabling Regularization Tuning

Xubo Yue, PhD Candidate, Department of Industrial & Operations Engineering
Co-Authors: Raed Al Kontar

We introduce an alternative closed form objective function for the Gaussian process (GP) based on the R'enyi alpha-divergence. This objective unifies both the known Nystr"om approximation and the exact GP. Its key advantage is the capability to control and tune the enforced regularization on the model and thus is a generalization of the traditional variational GP. From a theoretical perspective, we provide the convergence rate and risk bound for inference using our proposed objective. Experiments on real data show that the proposed method may deliver significant improvement over several GP inference techniques.

O1c. Analysis of Readmissions Data Taking Account of Competing Risks
Wenbo Wu, PhD Candidate, Department of Biostatistcs
Co-Authors: Kevin He, Xu Shi, Douglas E. Schaubel, John D. Kalbfleisch
Keyword: Discrete survival model; Cause-specific hazard; Provider profiling; Standardization; Robust inference

The 30-day unplanned hospital readmission rate has been used in healthcare provider profiling for evaluating inter-provider care coordination, medical cost effectiveness, and patient quality of life. Current profiling analyses use logistic regression to model readmission as a binary outcome, and the competing risks (e.g., death) and variable event times are not explicitly considered. Overlooking competing risks and event times leads to less comprehensive modeling and distorted provider evaluation. To address these drawbacks, we propose a discrete competing risk model within a cause-specific hazards framework. The discrete cause-specific hazard of readmission is used to assess provider-level effects. To facilitate the estimation of high-dimensional parameters, we develop a Blockwise Inversion Newton algorithm with scalability and memory efficiency. To draw inference about provider effects in the presence of patient-level repeated events, we devise a stabilized robust score test that overcomes the conservativeness of the classical robust score test, and proves suitable even for providers with extreme outcomes. Application to readmissions data from a national dialysis patient database demonstrates improved profiling, model fitting, and outlier detection over existing methods. Simulations mimicking the readmissions data display controlled type I error and enhanced power associated with the proposed test.

O1d. Image-on-Scalar Regression via Neural Networks
Daiwei Zhang, PhD Candidate, Department of Biostatistcs
Co-Authors: Lexin Li, Chandra Sripada, Jian Kang
Keyword: brain imaging, fMRI, functional data analysis, high-dimensional inference, model selection, neural networks

In medical imaging studies, a topic of central interest is the association analysis of massive imaging data with covariates of interest. The difficulty arises from the ultrahigh imaging dimensions, heterogeneous noises, and limited number of training images. To address these challenges, we propose a novel and conceptually straightforward neural network-based image-on-scalar regression model, in which the spatially varying functions of the main effects, individual deviations, and noise variances are all constructed through neural networks. Compared with existing methods, our method can identify a wider variety of spatial patterns, better captures the individual-wise heterogeneity, and is less affected by a small number of individuals. We provide estimation and selection algorithms with theoretically guaranteed asymptotic properties when the number of voxels grows faster than the number of individuals. We demonstrate the efficacy of our method through extensive simulation studies and the analysis of the fMRI data in the Autism Brain Imaging Data Exchange study and the Adolescent Brain Cognitive Development study.

O1e. Separating and integrating latent variables for improved classification of genomic data

Yujia Pan, PhD Candidate, Department of Statistics
Co-Authors: Johann A. Gagnon-Bartsch
Keyword: ensemble, gene expression, dimension reduction

Genomic datasets contain the effects of various unobserved biological variables in addition to the variable of primary interest. These latent variables often affect a large number of features (e.g., genes), giving rise to dense latent variation. This latent variation presents both challenges and opportunities for classification. While some of these latent variables may be partially correlated with the phenotype of interest and thus helpful, others may be uncorrelated and merely contribute additional noise. Moreover, whether potentially helpful or not, these latent variables may obscure weaker effects that impact only a small number of features but more directly capture the signal of primary interest. To address these challenges, we propose the cross-residualization classifier (CRC). Through an adjustment and ensemble procedure, the CRC estimates and residualizes out the latent variation, trains a classifier on the residuals, and then re-integrates the the latent variation in a final ensemble classifier. Thus, the latent variables are accounted for without discarding any potentially predictive information. We apply the method to simulated data and a variety of genomic datasets from multiple platforms. In general, we find that the CRC performs well relative to existing classifiers and sometimes offers substantial gains.

Oral Presentations II

O2a. Latent space models for multiplex networks with shared structure
Peter MacDonald, PhD Candidate, Department of Statistics
Co-Authors: Elizaveta Levina, Ji Zhu
Keywords: Networks, Multilayer, Multiplex, Latent space model

Latent space models are frequently used for modeling single-layer networks and include many popular special cases, such as the stochastic block model and the random dot product graph. However, they are not well-developed for more complex network structures, which are becoming increasingly common in practice. Here we propose a new latent space model for multiplex networks: multiple, heterogeneous networks observed on a shared node set. Multiplex networks can represent a network sample with shared node labels, a network evolving over time, or a network with multiple types of edges. The key feature of our model is that it learns from data how much of the network structure is shared between layers and pools information across layers as appropriate. We establish identifiability, develop a fitting procedure using convex optimization in combination with a nuclear norm penalty, and prove a guarantee of recovery for the latent positions as long as there is sufficient separation between the shared and the individual latent subspaces. We compare the model to competing methods in the literature on simulated networks and on a multiplex network describing the worldwide trade of agricultural products.

O2b. Space-Time Adaptive Detection at Low Sample Support
Robert P Malinas, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Benjamin D. Robinson, Alfred Hero
Keywords: covariance estimation, detection theory, high-dimensional asymptotics, random matrix theory, spiked covariance, shrinkage, adaptive matched filtering

An important problem in space-time adaptive detection is the estimation of the large p √ó p interference covariance matrix from training signals. When the number of training signals n is greater than 2p, existing estimators are generally considered to be adequate, as demonstrated by fixed-dimensional asymptotics. But in the low-sample-support regime (n < 2p or even n < p), fixed-dimensional asymptotics are no longer applicable. The remedy undertaken in this paper is to consider the ‚Äúlarge dimensional limit‚Äù in which n and p go to infinity together. In this asymptotic regime, we define a new consistency condition for the so-called ‚Äúshrinkage‚Äù covariance estimators, we show that an estimator from the literature satisfies this consistency condition, and we show that any estimator satisfying the consistency condition is asymptotically ideal in terms of detection rate. Further, asymptotic detection and false-alarm rates of filters formed from this type of estimator are characterized and shown to depend only on data that is given, even for non-Gaussian interference statistics. Finally, we present Monte-Carlo simulations illustrating these results and their empirical convergence rates on simulated radar data.

O2c. Scalar on Network Regression via Boosting
Emily Morris, PhD Candidate, Department of Biostatistics
Co-Authors: Jian Kang
Keywords: Brain imaging, network analysis, fMRI, boosting

Neuroimaging studies have a growing interest in learning the association between the individual brain connectivity networks and their clinical characteristics. It is also of great interest to identify the sub brain networks as biomarkers to predict the clinical symptoms, such as disease status, potentially providing insight on neuropathology. This motivates the need for developing a new type of regression model where the response variable is scalar, and predictors are networks that are typically represented as adjacent matrices or weighted adjacent matrices, to which we refer as scalar-on-network regression. In this work, we develop a new boosting method for model fitting with sub-network markers selection. Our approach, as opposed to group lasso or other existing regularization methods, is essentially a gradient descent algorithm leveraging known network structure. We demonstrate the utility of our methods via simulation studies and analysis of the resting-state fMRI data in a cognitive developmental cohort study.

O2d. Functional Data Analysis with Ultra High-Dimensional Covariates
Michael Law, PhD Candidate, Department of Statistics
Co-Authors: Ya'acov Ritov
Keywords: Brain imaging, network analysis, fMRI, boosting

We consider a sparse high-dimensional varying coefficients model with random effects, a flexible linear model allowing covariates and coefficients to have a functional dependence with time. For each individual, we observe discretely sampled responses and covariates as a function of time as well as time invariant covariates. Under sampling times that are either common or independent amongst individuals, we propose a projection procedure for the empirical estimation of all varying coefficients. We extend this estimator to construct confidence bands for a fixed number of varying coefficients.

O2e. A Bayesian stacked imputation framework for extended regression inference in data integration

Tian Gu, PhD Candidate, Department of Biostatistics
Co-Authors: Jeremy M.G. Taylor, Bhramar Mukherjee
Keywords: Data integration; Prediction models; Stacked multiple imputation

In the era of big data, it has become increasingly common for researchers to consider incorporating external information from large-scale studies to improve statistical inference rather than using the limited-sized data that are available to each investigator. However, challenges exist such as data sharing, storage, and privacy issues that limit access to the individual-level large data, but often it is easy to obtain the summary information. Therefore, there is a growing need for general frameworks that integrate the individual-level data and the summary-level external information in a principled manner. Aiming to fit a regression model for the outcome on all available covariates, we propose a unified framework that addresses the following challenges: (i) incorporating supplementary information from a broad class of externally fitted predictive models or established risk calculators, as long as the external model can generate outcome values given covariates; (ii) missing data imputation for using external models based on only a subset of all the covariates; (iii) improving statistical efficiency of the estimated coefficients in the internal study; and (iv) valid statistical inference for the external population with potentially different covariate effects from the internal study. Applications include prostate cancer risk prediction models using novel biomarkers that are measured only in the internal study.

Poster Session II

P2a. Parametric Models vs. Trees for Missing Data Imputation
Micha Fischer, PhD Candidate, Program in Survey Methodology
Keywords: Multiple Imputation, Model Selection, Missing Values, Sequential Imputation

Multiple sequential imputation (MSI) is a common way to deal with missing values in (survey) data. To carry out MSI, many different procedures have been proposed over the past two decades. Although many studies compared several methods for missing data imputation and have advanced the field, they often rely on one assessment strategy (e.g., simulated parametric data) only and often compare only a small number of procedures and model types (i.e. the current standard and the new proposed procedure). These shortcomings lead to findings with low generalizability. Since different methods (most likely) favor different data situations, this study compares a set of parametric models (Bayesian (regularized) linear models, predictive mean matching) with several tree-based models (CART, random forest, BART) for MSI of missing data in three ways: a comparison based on 1) parametric data; 2) nonparametric data; and 3) a real data set. The poster presentation will provide preliminary results based on parametric data.

P2b. Weston-Watkins Hinge Loss and Ordered Partitions
Yutong Wang, PhD Candidate, Department of Electrical Engineering and Computer Science
Co-Authors: Clayton Scott

Classification is a central problem in supervised learning, where the goal is to learn a decision function that accurately assigns labels to instances. The support vector machine (SVM) is a learning algorithm that is popular in practice and also has strong theoretical properties. However, most of the theory developed is for the binary classification setting, where there are only two possible labels to choose from. Our work is concerned with the multiclass setting where there are three or more possible labels for the decision function to choose from. Multiclass SVMs have been formulated in a variety of ways. A recent empirical study by Doğan et al. compared nine such formulations and recommended the variant proposed by Weston and Watkins (WW). Despite the superior empirical performance of the WW multiclass SVM, its theoretical properties remain poorly understood. Towards bridging this gap, we establish a connection between the hinge loss used in the WW multiclass SVM with ordered partitions. We use this connection to justify the recent empirical findings.

P2c. Expectation Propagation Linear Unmixing
Haonan Zhu, PhD Candidate, Department of Electrical Engineering and Computer Science
Co-Authors: Alfred Hero
Keywords: Expectation Propagation, Linear Unmixing, variational approximation

Expectation Propagation (EP) is a variational approximation algorithm that is efficient and scalable for high dimension problems. In this talk we show how it can be applied to a spectral unmixing problem arising in monitoring for nuclear safeguards. We discuss its convergence properties and demonstrate EP,Äôs empirical convergence for real world data. In addition, we show analytically that in the simple case Gaussian linear unmixing EP is equivalent to a quasi-newton algorithm.

P2d. Posterior inference for quantile regression -- adaptation to sparsity

Yuanzhi Li, PhD Candidate, Department of Statistics
Co-Authors: Xuming He
Keywords: Asymmetric Laplace distribution; Working likelihood; Shrinkage prior.

Quantile regression is a powerful data analysis tool that accommodates the heterogeneous relationship between a response variable and several covariates. In this talk, we examine the posterior inference for possibly sparse quantile regression models by coupling the asymmetric Laplace working likelihood with appropriate shrinkage priors.
With a feasible adjustment on the posterior variance, we find that the posterior inference is asymptotically valid and can automatically adapt to model sparsity, i.e., it achieves oracle efficiency for the active (non-zero) coefficients and super-efficiency for the inactive ones. The Bayesian computational framework demonstrates desirable inference stability due to its two distinct features: first, it avoids the need to pursue dichotomous variable selection; second, it circumvents direct estimation of unknown nuisance parameters.

P2e. Dimension Reduction and Nonparametric Multivariate Regression for Assessing Moderated Effects with Binary Outcome Trajectories

Xiru Lyu, Master student, Department of Statistics
Co-Authors: Galit Dunietz, Louise O,ÄôBrien, Ronald Chervin, Kerby Shedden
Keywords: dimension reduction, nonparametric multivariate regression, canonical correlation analysis, sleep-wake patterns, employment status, American Time Use Survey

We develop a nonparametric approach based on dimension reduction to model the conditional mean E[Y|X] where Y is a long binary trajectory and X is a vector of covariates. With basis expansion and canonical correlation analysis (CCA), we identify core nonlinear regression relationships between features derived from Y and from X to model E[Y|X] by local linear regression. By partitioning X into a key variable (X1), moderators (X2), and nuisance factors (X3), we can perform hypothesis tests based on log risk ratios, log odds and probability differences to assess the mean effect of X1 with fixed X2 values averaged over X3 by matching in the CCA space. We apply the technique to American Time Use Survey data to model associations between employment status and sleep-wake patterns, with individual sleep-wake trajectories of 1440 minutes as Y, employment status as X1 and demographics as X2, while controlling for secular trends (X3). Our findings highlight on average later bedtime and wake-time for non-employed vs. employed adults, with more pronounced effects in racial minorities. Results also suggest later wake-time for people of lower educational level when they become unemployed.

P2f. Evaluating the effect of ACA Medicaid expansion on mortality: Connecting ethnographic and statistical analyses through penalized matching

Charlotte Z Mann, PhD Pre-candidate, Department of Statistics
Co-Authors: Ben Hansen and Lauren Gaydosh
Keywords: Affordable care act, Medicaid, optimal matching, full matching.

States are able to choose whether to expand Medicaid as part of the Affordable Care Act (ACA); thus it is of interest to understand the impact of this policy choice on the health of residents in each state. To estimate the causal effects of Medicaid expansion on mortality in the US, we optimally matched counties from expanding and non-expanding states. To benefit from separate ethnographic studies which investigate why white individuals might oppose such policy when Medicaid expansion may benefit them directly (Metzl, 2019), our matching procedure optimized a criterion including but not limited to differences on a propensity score.
While maintaining a quality match in terms of covariate balance between the treatment and control counties as groups, we introduced penalties to the matching distance to ensure that similarly close specific counties would also be paired. Our procedure furnished closely-matched pairs of counties from all over the country, some mutually adjacent and others geographically separated. This aspect of the structure sets the stage for future ethnographic work (i.e. focus groups and interviews in these matched counties). Reviewing these matches closely can also reinforce the validity of the matching structure, or suggest methods for improvement.

P2g. A Beta Splitting Model for Coalescent Trees

Enes Dilber, PhD Candidate, Department of Statistics
Co-Authors: Jonathan Terhorst
Keywords: Population Genetics, Natural Selection, Genetics

Understanding evolution and natural selection is a long-standing scientific challenge. In this work, we propose new statistical methods for studying natural selection. Selection distorts genealogies in ways that are observable given sufficiently large samples of genetic data. We derive new methods for detecting this signal. Our main contribution is a new parametric model, based on the idea of Aldous (1996), for generating imbalanced coalescent trees. By fitting this model to data, we can determine genes that are affected by natural selection, as well as the particular form (e.g., directional or balancing) of selection that is acting. Compared to existing methods in this space, which are mostly likelihood-free, ours is derived from an explicit probabilistic model of tree imbalance. This conveys certain advantages for estimation and testing. We generalize Aldous' so-called beta-splitting model to accommodate dependence over time, and provide several methods for fitting the model depending on the type of data one has available. We test our model performance on the 1000 Genomes Project. In our preliminary work, we showed our model works as expected in the regions known to experience selection.

P2h. General-Sum Markov Games with Piecewise Stationary Opponent Policies
Anthony DiGiovanni, Pre-candidate, Department of Statistics
Keywords: reinforcement learning, game theory, change detection, learning theory

Reinforcement learning problems with multiple agents pose the challenge of ensuring that a learning agent's policy adapts efficiently to nonstationary dynamics, which arise from the strategic behavior of the other agents. Although a number of algorithms have been designed for these problems with promising empirical results, regret analysis of such algorithms in general-sum games has been very limited. I propose an algorithm for general-sum Markov games against opponents that follow a sequence of stationary policies (TSMG), which combines change detection with Thompson sampling to quickly learn a sequence of parametric models of the opponent. I prove that under standard assumptions for parametric MDP learning and given L-1 sufficiently infrequent policy switches by the opponent, the expected regret of TSMG in the worst case over switch schedules and opponent parameters is $O(L + L^{1/2}T^{1/2(log(Tlog(T)))^{1/2}})$. I validate this regret bound with experiments in several classic games, showing that TSMG can outperform both standard Thompson sampling model-based RL and a parametric version of UCRL2 with resets, even when the assumption of ergodicity is violated.

P2i. Bayesian Inference for Brain Activity from Functional Magnetic Resonance Imaging Collected at Two Spatial Resolutions
Andrew Whiteman, PhD Candidate, Department of Biostatistics
Co-Authors: Andreas J. Bartsch, Jian Kang, Timothy D. Johnson
Keywords: Bayesian Nonparametrics, Medical Imaging, Gaussian Process, Data Integration

Neuroradiologists and neurosurgeons may opt to use functional magnetic resonance imaging (fMRI) to map functional brain regions and plan out surgical access routes noninvasively. This application requires a high degree of spatial accuracy, but the fMRI signal-to-noise ratio (SNR) decreases as spatial resolution increases. In practice, fMRI scans can be collected at multiple spatial resolutions, and it is of interest to make more accurate inference on brain activity by combining data with different resolutions. To this end, we develop a new Bayesian model to leverage both better spatial precision in high resolution fMRI and higher SNR in standard resolution fMRI. We assign a Gaussian process prior to the mean intensity function and develop an efficient, scalable posterior computation algorithm to integrate both sources of data. We draw posterior samples using an algorithm analogous to Riemann manifold Hamiltonian Monte Carlo in an expanded parameter space. We illustrate our method in analysis of presurgical fMRI data, and show in simulation that it infers the mean intensity more accurately than alternatives that use either the high or standard resolution fMRI data alone.

Oral Presentations III

O3a. A Multivariate Stopping Rule for Survey Data Collection
Xinyu Zhang, PhD Pre-candidate, Program in Survey Methodology
Co-Authors: James Wagner, Michael R. Elliott, Brady T. West, and Stephanie Coffey
Keywords: Stopping rule; responsive design; nonresponse bias

Surveys are experiencing declining response rates. With more and more effort expended to combat these declining response rates, the cost of survey data collection has continued to rise. Recent technological developments in survey data collection have allowed the survey designer to intervene early, possibly stopping effort on selected active cases that are not expected to increase quality sufficiently given their projected costs. In multipurpose surveys, there may be data quality objectives that must be met for certain estimates with constraints on costs. We introduce a stopping rule for survey data collection that accounts for the quality of more than one survey variable. The proposed stopping rule is illustrated via simulation using data from the Health and Retirement Study.

O3b. Estimation of Knots in Linear Spline Model
Guangyu Yang, PhD Candidate, Department of Biostatistics
Co-Authors: Baqun Zhang, Min Zhang
Keywords: Broken-stick Model; Change-Point; Efficiency; Semiparametric Theory; Threshold Effect

The linear spline model accommodates nonlinear effects while maintaining easy interpretation. It has significant applications in studying threshold effects and change-points. However, its application in practice has been limited by the lack of rigorously studied and computationally convenient method for estimating knots. A key difficulty in estimating knots lies in nondifferentiability. In this study, we study influence functions of regular and asymptotically linear estimators for linear spline models using semiparametric theory rigorously. Based on the theoretical development, we propose a novel and simple method to circumvent the nondifferentiability using modified derivatives, in contrast to previous smoothing-based methods. Consistency and asymptotic normality are rigorously derived for the estimator using empirical process theory. To improve numerical stability, a two-step algorithm taking advantage of the analytic solution available when knots are known is developed. Simulation studies have shown the two-step algorithm performs well in terms of both statistical and computational properties and offers a substantial improvement over existing methods.

O3c. A Multi-batch L-BFGS Method with Variance Reduction: Theory and Experiments
Zihong Yi, Undergraduate student, Department of Industrial & Operations Engineering
Co-Authors: Albert S. Berahas
Keywords: Nonlinear Optimization; Machine Learning; Stochastic Gradient; Variance Reduction; Quasi-Newton

In this talk, we present a new class of stochastic quasi-Newton methods for solving optimization problems that arise in machine learning. The class of methods constructs a gradient approximation using a fraction of the data, leverages the BFGS updating formula to construct curvature information, and utilizes variance reduction techniques to mitigate the inherent variance that arise in the stochastic setting. Specifically, we combine the multi-batch LBFGS method with two state-of-the-art variance reduction techniques (SVRG and SARAH). We prove theoretical convergence guarantees for strongly convex and nonconvex problems, and illustrate the performance of the method on standard machine learning training tasks.

O3d. Inference post Selection of Group-sparse Regression Models
Daniel Kessler, PhD Candidate, Department of Statistics
Co-Authors: Snigdha Panigrahi, Peter W. MacDonald
Keywords: Conditional inference, Group-sparse, Linear Models, Randomization, Selective inference

Conditional inference provides a rigorous approach to counter bias when data from automated model selections is reused for inference. We develop in this paper a statistically consistent Bayesian framework to assess uncertainties within linear models that are informed by grouped sparsities in covariates. Finding wide applications when genes, proteins, genetic variants, neuroimaging measurements are grouped respectively by their biological pathways, molecular functions, regulatory regions, cognitive roles, these models are selected through a useful class of group-sparse learning algorithms. An adjustment factor to account precisely for the selection of promising groups, deployed with a generalized version of Laplace-type approximations is the centerpiece of our new methods. Accommodating well known group-sparse models such as those selected by the Group LASSO, the overlapping Group LASSO, the sparse Group LASSO etc., we illustrate the efficacy of our methodology in extensive experiments and on data from a human neuroimaging application.

O3e. ArgoSSM: A Bayesian state-space framework for predicting the location of oceanographic sensors in the Southern Ocean

Drew Yarger, PhD Candidate, Department of Statistics
Co-Authors: Derek Hansen
Keywords: Bayesian state-space model, particle filtering, oceanography, spatial statistics

The Argo project deploys a fleet of sensors that collect information such as the temperature and salinity at varying depths of the ocean. These sensors are attached to floats that drift with the ocean currents. In the Southern Ocean, these floats occasionally end up under ice, and their location can no longer be tracked via GPS. We introduce a novel framework, called ArgoSSM, to predict the location of these floats while they are under ice-cover. ArgoSSM is a fully probabilistic Bayesian state-space model which provides both point estimates and uncertainty in the missing location measurements. Moreover, it can incorporate additional information like potential vorticity in the predicted locations. We compare our approach to existing approaches in the oceanographic literature, such as linear interpolation. By providing a posterior distribution of potential paths the floats could have taken under ice, our modelled predictions and uncertainty can improve downstream tasks like temperature and salinity estimation, allowing for better scientific understanding of the Southern Ocean.

Notes