Michigan Student Symposium for Interdisciplinary Statistical Sciences

MSSISS 2019

March 28th, 2019

3:00pm – 6:30pm

Rackham Graduate School

March 29th, 2019

8:30am – 5:00pm

Palmer Commons

Sponsored by Biostatistics, EECS, IOE, MIDAS, Statistics, & Survey Methodology

Table of Contents

Committee & Acknowledgments

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

MSSISS 2019 Student Organizing Committee:

Tian Gu (Department of Biostatistics)
Julian Katz-Samuels (Department of Electrical Engineering & Computer Science)
Weiyu Li (Department of Industrial & Operations Engineering)
Rayleigh Lei (Department of Statistics)
Ai Rene Ong (Program in Survey Methodology)

MSSISS 2019 Faculty Advisory Committee:

Dr. Timothy Johnson (Department of Biostatistics)
Dr. Clayton Scott (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 insights. We thank Wendy Washburn and Jamie Clay of the Biostatistics Department for their help in organizing the event.

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

We are grateful to Anne Cain-Nielsen (MS), Nicholas Moloci (MPH), Karen E. Nielsen (PhD), Phyllis Yan (MS) and Ziwei Zhu (MS) of 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 2019

Official Schedule: Thursday, March 28th
Rackham Graduate Building, 4th floor

3:00 – 3:30pm **Registration** (East Conference Room)

3:30 – 3:35pm **Welcoming Remarks** (Amphitheater)

3:35 – 4:30pm **Speed Oral Presentations** (Amphitheater)

Timothy NeCamp, Department of Statistics
Developing year-long mobile health interventions to improve health outcomes among medical interns: experimental design and statistical methods

Yiwang Zhou, Department of Biostatistics
Net Benefit Index: A New Method of Biomarker Assessment in the Learning of Individualized Treatment Rules

Byoungwook Jang, Department of Statistics
Minimum Volume Topic Modeling

Yan-Cheng Chao, Department of Biostatistics
A Bayesian Dropping Rule for Small n Sequential Multiple Assignment Randomized Trial

Stephen Salerno, Department of Biostatistics
A Bayesian Factor Model for Healthcare Rankings: Applications in Estimating Composite Measures of Quality

Aditya Modi, Department of EECS
Contextual Markov Decision Processes using Generalized Linear Models

Derek Hansen, Department of Statistics
A Randomized Missing Data Approach to Robust Filtering with Applications to Economics and Finance

Aritra Guha, Department of Statistics
On posterior contraction of parameters and interpretability in Bayesian mixture modeling

Yutong Wang, Department of EECS
Unsupervised feature selection for manifold alignment of scRNA-seq data

Ali Rafei, Program of Survey Methodology
Bayesian Doubly Robust Adjustment for Finite Population Inference using Big Data: Application to Naturalistic Driving Studies

Michael Law, Department of Statistics
Inference Without Compatibility

4:30 – 5:30pm **Speed Presentation Poster Session, Interdisciplinary Poster Session, Appetizers** (East Conference Room)

5:30 – 6:25pm **Michigan Junior Faculty Keynote** (Amphitheater)

Assistant Professor Ceren Budak, University of Michigan
What happened? The Spread of Fake News Publisher Content During the 2016 Election

6:25 – 6:30pm **Closing Remarks** (Amphitheater)

MSSISS 2019

Official Schedule: Friday, March 29th
Palmer Commons, 4th floor

8:30 – 8:55am **Registration** (4th Atrium 4)**, Breakfast** (Great Lakes)

8:55 – 9:00am **Welcoming Remarks** (Forum Hall)

9:00 – 10:00am **Oral Presentations I** (Forum Hall)

Zhangchen Zhao, Department of Biostatistics

Methods to Account for Uncertainty in Latent Class Assignments when using Latent Classes as Predictors in Regression Models, with Application to Acculturation Strategy Measures

Xubo Yue, Department of IOE

Variational Inference of Joint Models using Multivariate Gaussian Convolution Processes

Fernanda Alvarado-Leiton, Program in Survey Methodology

Effects of statistical adjustments for subgroup differences in non-probability sample web surveys

10:00 – 11:15am **Poster Session I, Refreshments** (Great Lakes)

11:15 – 12:15pm **Oral Presentations II** (Forum Hall)

Jack Goetz, Department of Statistics

Active Learning for Nonparametric Regression Using Purely Random Trees

Seokhyun Chung, Department of IOE

Functional Principal Component Analysis for Extrapolating Multi-stream Longitudinal Data

Morteza Noshad, Department of EECS

Scalable Information Estimation for Deep Neural Networks

12:15 – 1:30pm **Lunch** (Great Lakes)

1:30 – 2:30pm **Oral Presentations III** (Forum Hall)

Pedro Orozco del Pino, Department of Biostatics

Use of simulations to study the impact of linkage disequilibrium in the distribution of genetic risk scores across populations

Elizabeth Hou, Department of EECS

Anomaly Detection in Partially Observed Traffic Networks

Baekjin Kim, Department of Statistics

On the Optimality of Perturbations in Stochastic and Adversarial Multi-armed Bandit Problems

2:30 – 3:45pm **Poster Session II**, **Coffee Break** (Great Lakes)

3:45 – 4:00pm **Student Awards** (Forum Hall)

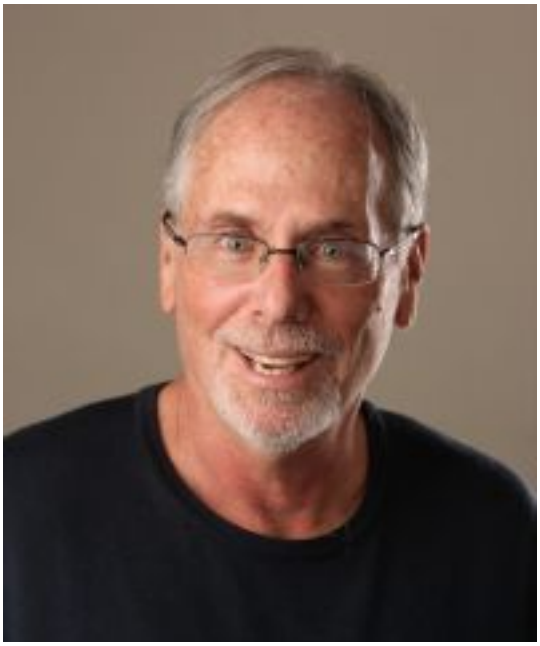
4:00 – 5:00pm **Keynote Address** (Forum Hall)

Professor Alan E. Gelfand, Duke University

Space is the Place: Why spatial thinking matters for environmental problems

5:00 – 5:15pm **Closing Remarks** (Forum Hall)

Keynote Speaker, Friday: Dr. Alan E. Gelfand



Alan E. Gelfand is the James B Duke Professor of Statistical Science at Duke University. He is the former chair of the Department of Statistical Science (DSS) and enjoys a secondary appointment as Professor of Environmental Science and Policy in the Nicholas School. Author of more than 290 papers (more than 230 since 1990), Gelfand is internationally known for his contributions to applied statistics, Bayesian computation and Bayesian inference. (An article in Science Watch found him to be the tenth most cited mathematical scientist in the world over the period 1991-2001). Gelfand is an Elected Fellow of the American Statistical Association, the Institute of Mathematical Statistics, and the International Society for Bayesian Analysis. He is an Elected Member of the International Statistical Institute. He is a former President of the International Society for Bayesian Analysis and in 2006 he received the Parzen Prize for a lifetime of research contribution to Statistics. In 2012, he was chosen to give the distinguished Mahalanobis lectures. In 2013, he received a Distinguished Achievement Medal from the ASA Section on Statistics in the Environment.

Gelfand's primary research focus for the past twenty years has been in the area of statistical modeling for spatial and space-time data. Through a collection of more than 140 papers he has the advanced methodology, using the Bayesian paradigm, to associate fully model-based inference with spatial and space-time data displays. His chief areas of application include environmental exposure, spatiotemporal ecological processes, and climate dynamics. He has four books in this area, including the successful "Hierarchical Modeling and Analysis for Spatial Data" with Sudipto Banerjee and Brad Carlin (now second edition), "Hierarchical Modeling for Environmental Data; Some Applications and Perspectives" with James Clark, and the "Handbook of Spatial Statistics" with Peter Diggle, Montserrat Fuentes, and Peter Guttorp. He also has a forthcoming "Handbook of Environmental and Ecological Statistics" with Montserrat Fuentes, Jennifer Hoeting, and Richard Smith.

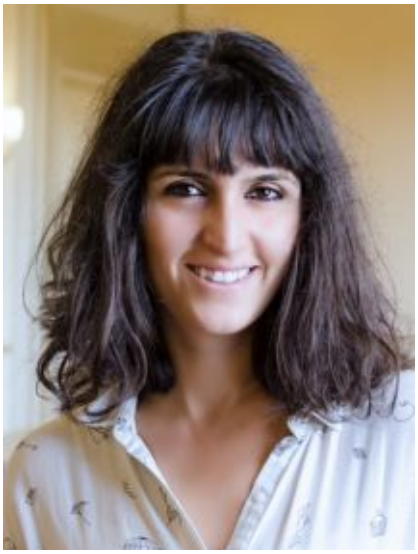
Title: Space is the Place: Why spatial thinking matters for environmental problems

Spatial methods have become an increasingly used approach for analyzing data in many fields. In particular, it is now routine to collect data layers where there is some geographic referencing. This information should be used in order to enhance inference. From a statistical perspective, we think in terms of formal inference, utilizing probabilistic or stochastic modeling; we think beyond purely descriptive summaries. In this sense, we exceed the capabilities of Geographic Information Systems (GIS) software to investigate complex processes over space and time.

A particularly rich context for such investigation is environmental processes. Examples include analysis of weather/climate data, analysis of environmental exposure data, analysis of locations of disease occurrence, and analysis of distributions of species over a region. In this non-technical talk, I will describe the types of spatial (and, perhaps, spatio-temporal) data that we collect. I will discuss what we expect to see with regard to these types of data, i.e., what we mean by "spatial pattern." I will raise a variety of issues that arise in modeling such data - explanation of local behavior through spatially referenced explanatory variables, explanation of uncertainty through structured dependence. I will illustrate, with a variety of datasets involving the foregoing processes, hopefully to illuminate that statistical thinking does matter when we have inferential objectives such as explanation, interpolation, and prediction.

Finally, with increasingly larger monitoring networks and remote sensing from satellites, we are seeing data collected at continental or global scales, over space and time. Data fusion and "big data" challenges emerge and efficient model fitting strategies are needed. I will offer some thoughts in this context.

Michigan Junior Faculty Speaker, Thursday: Dr. Ceren Budak



Ceren Budak is an Assistant Professor at the University of Michigan School of Information. Before that, she was a Postdoctoral Researcher at Microsoft Research New York. She received her Ph.D. from the Computer Science Department at the University of California, Santa Barbara in December 2012. She received her Bachelor's degree from Computer Science Dept. @Bilkent University in Turkey in 2007.

Her research interests lie in the area of computational social science; a discipline at the intersection of computer science, statistics, and the social sciences. She is particularly interested in applying large-scale data analysis techniques to study problems with social, political, and policy implications.

Title: What happened? The Spread of Fake News Publisher Content During the 2016 Election

The spread of fake news was one of the most discussed characteristics of the 2016 U.S. Presidential Election.The concerns regarding fake news have garnered significant attention in both media and policy circles, with some journalists even going as far as claiming that results of the 2016 election were a consequence of the spread of fake news. Yet, little is known about the prevalence and focus of such content, how its prevalence changed over time, and how this prevalence related to important election dynamics. In this talk, I will address these questions by examining social media, news media, and interview data. These datasets allow examining the interplay between news media production and consumption, social media behavior, and the information the electorate retained about the presidential candidates leading up to the election.

Speed Oral Presentations

S1. Developing Year-long Mobile Health Interventions to Improve Health Outcomes Among Medical Interns: Experimental Design and Statistical Methods
Timothy NeCamp, PhD Candidate, Department of Statistics
Co-Authors: Zhenke Wu, Srijan Sen
Keywords: Mobile health, Micro-randomized trial, Experimental design

Medical interns tend to work long hours, undergo stress, sleep inconsistently, face difficult decisions, and cope with mental health issues during their internship year. There is a critical need to develop interventions to help these interns improve their mood, stay physically active, and sleep consistently. Mobile health interventions (interventions delivered on a phone or mobile device) hold promise because interventions can be delivered any time at low burden to interns. Unfortunately many questions regarding the design, delivery, and efficacy of these mobile health interventions remain unanswered.

S2. Net Benet Index: A New Method of Biomarker Assessment in the Learning of Individualized Treatment Rules
Yiwang Zhou, PhD Candidate, Department of Biostatistics
Co-Authors: Haoda Fu, Peter X.K. Song
Keywords: Bootstrap; Clinical trial; O-learning; Personalized medicine; Variable selection

One central task in the personalized medicine paradigm is to establish individualized treatment rules (ITRs) for patients with heterogeneous responses to different therapies. Motivated from a diabetes clinical trial, we consider a problem where there exists a set of candidate biomarkers potentially useful to improve an existing ITR in a current treatment protocol. This calls for a biomarker assessing procedure that enables to evaluate the added values of the individual biomarkers. We put forth a screening analytic term as net benefit index (NBI) that quantifies the contrast between gain and loss of treatment benefits due to reallocation of patients in the treatment arms. The optimal treatment group labels are determined by support vector machine (SVM) under the context of outcome weighted learning (OWL). We propose an NBI-based test for significance of a biomarker improving the existing ITR with the bootstrap null distribution generated by stratified permutation within each treatment arm. Analytic results shows that baseline fasting insulin is the only variable that significantly improves the existing ITR involving age, body mass index (BMI) and baseline fasting plasma glucose (FPG) in the assignment of pioglitazone and gliclazide for patients with Type 2 diabetes with respect to the average reducing rate of FPG during the 52 weeks of treatment.

S3. Minimum Volume Topic Modeling
Byoungwook Jang, PhD Candidate, Department of Statistics
Co-Authors: Alfred Hero

We propose a new topic modeling procedure that takes advantage of the fact that the Latent Dirichlet Allocation (LDA) log-likelihood function is asymptotically equivalent to the logarithm of the volume of the topic simplex. This allows topic modeling to be reformulated as finding the probability simplex that minimizes its volume and encloses the documents that are represented as distributions over words. A convex relaxation of the minimum volume topic model optimization is proposed, and it is shown that the relaxed problem has the same global minimum as the original problem under the separability assumption and the sufficiently scattered assumption introduced by Arora et al. (2013) and Huang et al (2016). A locally convergent alternating direction method of multipliers (ADMM) approach is introduced for solving the relaxed minimum volume problem. Numerical experiments illustrate the benefits of our approach in terms of computation time and topic recovery performance.

S4. A Bayesian dropping rule for small n sequential multiple assignment randomized trial
Yan-Cheng Chao, PhD Candidate, Department of Biostatistics
Keywords: Interim Analysis, Type I error rate

A sequential multiple assignment randomized trial (SMART) is a multi-stage design where subjects may be re-randomized to treatment based on intermediate endpoints. In a standard SMART that we are interested in, subjects are randomly assigned to one of the three treatment options in the first stage. Participants who respond to first stage treatment, continue with their initial treatment, while participants who do not respond are randomized to one of the other two treatments that they did not initially receive. However, it is possible that if some subjects fail to respond to the initial treatment, they will be randomized to a treatment in the second stage that later is found to be ineffective. To mitigate this problem, but still provide valid inference for the treatment effects, we would like to incorporate interim analyses into the SMART design. By formulating an interim decision rule for dropping one of the treatments, we use Bayesian methods and the resulting posterior distributions to provide sufficient evidence that one treatment is inferior to the other treatments before enrolling more subjects. By doing so, we increase the likelihood that fewer subjects are assigned to the inferior treatment. Based on simulation results, we have evidence that the treatment response rates can be unbiasedly estimated for the better treatments in our new design. In addition, by adjusting the decision rule criteria for the posterior probabilities, we can control the Type I error rate of incorrectly dropping an effective treatment.

S5. A Bayesian Factor Model for Healthcare Rankings: Applications in Estimating Composite Measures of Quality

Stephen Salerno, PhD Pre-candidate, Department of Biostatistics
Co-Authors: Stephen Salerno, MS; Lili Zhao, PhD; and Yi Li, PhD
Keywords: Latent Variable Methods, Bayesian

Due to increased public interest in the reporting of healthcare metrics, national quality initiatives have focused on greater transparency and interpretability. Ranking statistics are of vital importance to public reporting efforts as they provide an intuitive means of conveying information to patients. Rankings are often based on composite measures, which aggregate information from multiple complex metrics. As these individual measures arise from different underlying distributions, with different scales, and missing observations, computation becomes particularly challenging. We offer a flexible, rank-based Bayesian factor model for the estimation of such composite observation ranks. We employ the parsimony of Bayesian MCMC for semi-parametric copula estimation and simultaneous missing imputation while resolving identifiability issues with factor rotation. Through this method, the factor scores can be viewed as a scale-free surrogate for estimating the underlying order statistics of our joint distribution. We show in simulation and a real data example that this method can outperform other ad-hoc deterministic or parametric model-based aggregation approaches in certain settings.

S6. Contextual Markov Decision Processes using Generalized Linear Models

Aditya Modi, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Ambuj Tewari
Keywords: Reinforcement Learning, Regret Bounds, Generalized Linear Models, Personalized Learning

We consider the recently proposed reinforcement learning (RL) framework of Contextual Markov Decision Processes (CMDP), where the agent interacts with infinitely many tabular environments in a sequence. In this paper, we propose a no-regret online RL algorithm in the setting where the MDP parameters are obtained from the context using generalized linear models (GLM). The proposed algorithm GL-ORL is completely online and memory efficient and also improves over the known regret bounds in the linear case. In addition to an Online Newton Step based method, we also extend existing tools to show conversion from any online no-regret algorithm to confidence sets in the multinomial GLM case. We also provide experimental results on toy domains to verify our guarantees.

S7. A Randomized Missing Data Approach to Robust Filtering with Applications to Economics and Finance

Derek Hansen, PhD Pre-candidate, Department of Statistics
Co-authors: Dobrislav Dobrev, Pawel Szerszen
Keywords: State-Space Models, Particle Filter, Inflation Forecasting, Stochastic Volatility, Realized Volatility, Outliers, Robust Estimation

We put forward a simple new approach to robust filtering of state-space models, motivated by the idea that the inclusion of only a small fraction of available highly precise measurements can still extract most of the attainable efficiency gains for filtering latent states, estimating model parameters, and producing out-of-sample forecasts. The new class of particle filters we develop aims to achieve a degree of robustness to outliers and model misspecification by purposely randomizing the subset of utilized highly precise but possibly misspecified or outlier contaminated data measurements, while treating the rest as if missing. The arising robustness-efficiency trade-off is controlled by varying the fraction of randomly utilized measurements or the incurred relative efficiency loss from such randomized utilization of the available measurements. As an empirical illustration, we consider popular state space models for inflation and equity returns with stochastic volatility and document favorable performance of our robust particle filter and density forecasts on both simulated and real data. More generally, our randomization approach makes it easy to robustly incorporate highly informative but possibly contaminated modern "big data" streams for improved state-space filtering and forecasting.

S8. On posterior contraction of parameters and interpretability in Bayesian mixture modeling

Aritra Guha, PhD Candidate, Department of Statistics
Co-Authors: Nhat Ho, XuanLong Nguyen
Keywords: Mixture models, Bayesian nonparametrics, Misspecification, Posterior consistency

We study posterior contraction behaviors for parameters of interest in the context of Bayesian mixture modeling, where the number of mixing components is unknown while the model itself may or may not be correctly specified. Two representative types of prior specification will be considered: one requires explicitly a prior distribution on the number of mixture components, while the other places a nonparametric prior on the space of mixing distributions. The former is shown to yield an optimal rate of posterior contraction on the model parameters under minimal conditions, while the latter can be utilized to consistently recover the unknown number of mixture components, with the help of a fast probabilistic post-processing procedure. We then turn the study of these Bayesian procedures to the realistic settings of model misspecification. It will be shown that the modeling choice of kernel density functions plays perhaps the most impactful roles in determining the posterior contraction rates in the misspecified situations. Drawing on concrete posterior contraction rates established in this paper we wish to highlight some aspects about the interesting tradeoffs between model expressiveness and interpretability that a statistical modeler must negotiate in the rich world of mixture modeling.

S9. Unsupervised feature selection for manifold alignment of scRNA-seq data
Yutong Wang, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Tasha Thong, Justin Colacino, Venkatesh Saligrama, Laura Balzano, Clayton Scott
Keywords: Single-cell RNA-seq, data integration

For integrating multiple single-cell RNA-sequencing datasets, a crucial step is to select a subset of relevant genes for the downstream alignment algorithm. Despite the importance of alignment, no work has been done on how to identify genes that have high alignment potential. We present a method for ranking the relevance of genes with respect to this task. Our method is unsupervised, i.e., the cell-types are not assumed to be known. Using the top-ranking genes significantly improves the downstream performance of alignment algorithms measured in terms of batch mixing and preserving known labels. We demonstrate the effectiveness of our algorithm on pre-implantation embryos development, neurogenesis, liver bud development, and mammary gland morphogenesis datasets.

S10. Bayesian Doubly Robust Adjustment for Finite Population Inference using Big Data: Application to Naturalistic Driving Studies
Ali Rafei, PhD Pre-candidate, Program in Survey Methodology
Co-Authors: Michael R. Elliott, Carol A.C. Flannagan
Keywords: partially observed Markov process models; state space models; particle filter; curse of dimensionality

With the widespread availability of Big Data, concerns are raised over finite population inference based on such large-scale non-probability samples. In presence of a benchmark survey with relevant auxiliary variables, one might apply a doubly robust adjustment by combining pseudo-weights with a prediction model to further protect against model misspecification. Traditionally, inverse propensity scores are used as pseudo-weights, but this method lacks adequate justification when auxiliary variables are partially observed. We propose a theoretically valid alternative approach to augment the prediction model in non-probability sample settings. Since the true model is often unknown, and Big Data tend to be poor in such model-relevant covariates, we employ Bayesian additive regression trees, which provide a flexible nonparametric predictive tool. In addition, a bootstrap method is adopted to incorporate the uncertainty in both pseudo-weights and outcome variable into variance estimation. Considering the National Household Travel Survey 2017 as benchmark, we apply our method to improve the generalizability of naturalistic driving data in the Strategic Highway Research Program 2.

S11. Inference Without Compatibility
Michael Law, PhD Candidate, Department of Statistics

We consider hypothesis testing problems for a single covariate in the context of a linear model with Gaussian design when p>n. Under minimal sparsity conditions of their type and without any compatibility condition, we construct an asymptotically Gaussian estimator with variance equal to the oracle least-squares. The estimator is based on a weighted average of all models of a given sparsity level in the spirit of exponential weighting. We adapt this procedure to estimate the signal strength and provide a few applications. We support our results using numerical simulations based on algorithm which approximates the theoretical estimator and provide a comparison with the de-biased lasso.

Interdisciplinary Poster Session

I1. Estimating Partisan Associations Using Word Embeddings
Patrick Y. Wu, PhD Candidate, Department of Political Science
Keywords: natural language processing, machine learning, scaling, dimension reduction, partisanship, twitter, social media, word embeddings

Twitter is a rich source of material for the study of political phenomena. A natural question is how a variable covaries with the users' partisan associations. The state-of-the-art approaches of estimating these associations involve analyzing which partisan actors a user follows. But such approaches often make unrealistic assumptions about why a user chooses to follow someone and are difficult to update. Our method, instead, analyzes the userâ€™s description. Using doc2vec, we first obtain document vectors for each user description and word vectors for every unique word across the descriptions. We then compare these user descriptions to a selection of partisan keywords, such as "clinton," "trump," "democrat," "republican," etc. by calculating the cosine similarities between the document vectors and the word vectors for these keywords. These cosine similarities can be considered "imperfect" observations of some latent partisanship trait, which can be recovered using techniques such as multidimensional scaling. We apply this technique to tweets that make some observation about the electoral process in the 2016 election in order to study how observations vary by partisan stances.

I2. The Tie that Binds: A Grouped Approach to Ideal Point Estimation
Kevin McAlister, PhD Candidate, Department of Political Science
Keywords: Bayesian Nonparametrics, Ideal Point Estimation, Generative Clustering, Time Series

Roll call scaling techniques are empirical standards for studies of voting behavior within legislative bodies. Though ideal point estimation techniques are frequently used, the theoretical implications of assumptions made in order to empirically estimate ideal points provide cause for concern. Current scaling techniques ignore the role of group-level dependencies within the data and this leads to potential biases in the estimated values of the ideal points. I propose a new ideal point model that allows for group contributions in the underlying spatial model of voting. I derive a corresponding empirical model that incorporates flexible Bayesian nonparametric priors to estimate group effects in ideal points and the corresponding dimensionality of the ideal points. I apply this model to the entire history of the U.S. House and show how group dynamics can be uncovered using only a set of roll call votes. Using this data, I explore the evolution of group voting within the U.S. legislative body. This model provides insights into open questions related to group dynamics in legislative voting and has important implications for literature that utilizes ideal point estimates.

I3. Recovering low-rank structure from multiple networks with unknown edge distributions

Keith Levin, Post-doc, Department of Statistics
Co-Authors: Asad Lodhia, Elizaveta Levina
Keywords: networks, low-rank estimation

In increasingly many settings, particularly in neuroscience, data sets consist of multiple samples from a population of networks, in which a notion of vertex correspondence across networks is present. For example, in the case of neuroimaging data, fMRI data yields graphs whose vertices correspond to brain regions that are common across subjects. The behavior of these vertices can thus be sensibly compared across graphs. We consider the problem of estimating parameters of the network population distribution under this setting. In particular, we consider the case where the observed networks share a low-rank structure, but may differ in the noise structure on their edges. Our approach exploits this shared low-rank structure to denoise edge-level measurements of the observed networks and estimate population-level parameters. We also explore the extent to which complexity of the edge-level error structure influences estimation and downstream inference.

I4. An Algorithm for Adjusted Kernel Linear Discriminant Analysis

Lynn Huang, Undergraduate, Department of Statistics
Keywords: Dimension reduction, Kernel Linear Discriminant Analysis, Classification, Facial Imaging

The current implementation of KLDA in R fails to compute projections in the case that the kernel matrix is non-invertible in the objective function. In this presentation, we propose an algorithm for adjusted KLDA which allows for the approximation of singular matrices within KLDA's objective function, ensuring the success of computations for any set of tuning parameters. The validity of the algorithm is evaluated on several simulated datasets, then applied to three versions of a subset of the Morph-II dataset containing different extracted features for face imaging tasks. The transformed feature set is used to train several statistical classification models, whose performance is then evaluated to determine the efficacy of the algorithm.

I5. Comparing Pycnophylactic and Kriging Methods Using Data Related to the Flint Water Crisis

Angelica Estrada, Post-Baccalaureate Alumna, Department of Statistics
Co-Authors: Zixian Li, Siyu Lily Qian, Sifan Jiang, Yi Wang, Nicole Keeping
Keywords: non-convex optimization, preference learning

This project focuses on the Modifiable Areal Unit Problem (MAUP) using data from the Flint Water Crisis in Michigan. MAUP is a problem in spatial statistics concerned with how different types of spatial aggregation used on the same data can lead to different results. The motivation of this project follows from the Flint Water Crisis, when officials reported statistics on the lead levels in the blood of individuals who lived within any of the ZIP codes of Flint. Since the ZIP code borders of Flint do not align with the city border of Flint, analyses included people who did not reside within the city of Flint and did not consume the water. Misrepresentation of the data may have produced biased, misleading results and serves as an example of how results are affected by the type of geographical or spatial unit used. The aim of this project is to transform the data from ZIP code level to the municipal level to allow for more accurate results. This report discusses using the kriging method to transform the data from one level of spatial aggregation to another in particular, from one polygon spatial type (ZIP code) to another polygon spatial type (city).

I6. Comparing Dynamic Treatment Regimes using Repeated-Measures Outcomes: An Iterative Estimation Method for SMART Studies

Madison Stoms, Undergraduate, Department of Statistics
Keywords: Dynamic Treatment Regime, SMART, Decision Making

Effective treatment of distinct patients warrants individualized and tailored interventions. A dynamic treatment regime (DTR) is a sequence of treatments adapted according to predetermined decision rules for an individual's specific needs. DTRs offer clinicians guidance on how to begin as well as change treatment over time, depending on the individualâ€™s response. Sequential multiple assignment randomized trials (SMART) are multi-stage, sequentially-randomized trial designs developed to inform the construction of a DTR. We consider repeated-measures SMART studies in which the primary goal is comparison between DTRs. Complications may arise when using over-the-counter statistical software to estimate parameters in these repeated-measures marginal models, due to the possibility of participants being randomized more than once, causing some individuals to be consistent with multiple regimes. Therefore, we propose a method to estimate these parameters by hand using an algorithm in R to iteratively estimate the covariance structure and regression coefficients. We demonstrate this method using data from a completed SMART study aimed at developing a DTR for increasing verbal expression in minimally verbal children on the Autism Spectrum.

I7. Optimising Dental Tool Dispensing at the University of Michigan School of Dentistry
Seungho Woo, Undergraduate, Department of Industrial & Operations Engineering
Co-Authors: Sejin Park
Keywords: innovation, industrial policy, venture capital, machine learning, network analysis

This paper documents the project to improve the University of Michigan School of Dentistry's dispensing and sanitation processes of dental equipment. The authors modelled the dispensing and sanitation process to find areas of inefficiencies and to develop solutions for improvement. With the modelling process, the authors chose to use ProModel 2018 as the main software of analysis. The layout of the sanitation process was then modeled with the entities, location, processing, variables etc. of ProModel. After the initial modelling process, areas of improvement, such as increasing sanitation staffing and batching dental equipment kits at different locations, were considered and then applied to the model. The new model was run and the results show that the efficiency of the system improved. The authors recommend implementing the alternative batching model to increase the availability of dental equipment kits and to decrease the workload of the sanitation workers.

I8. A novel temporal pathway analysis method to study the host transcriptomic response to influenza virus
Yaya Zhai, PhD Candidate, Department of Computational Medicine & Bioinformatics
Co-Authors: Yongsheng Huang
Keywords: influenza; viral infection; challenge study; temporal pathway analysis; graph Laplacian

Population could have divergent infection outcomes when they are challenged with the influenza virus, and some are more resistant than others. Previous study has discovered that the host response in asymptomatic subjects have a very distinct pattern from that in symptomatic subjects. However, the differences would be very difficult to interpret thus providing limited clinical perspective if they cannot be appropriately characterized by their biological functions. Here we propose a novel temporal pathway analysis method that combines graph Laplacian clustering with pathway significance measures to so that changes in multiple functionally-related genes over time could be taken into consideration to understand the complex and dynamic process in different response to influenza viral inoculation.

I9. Towards Question Recommendation: an Analysis of Intelligent Tutoring System Data
Jack Finkel, Laura Niss & Haonan Sun, Undergraduate, Department of Mathematics & Statistics
Keywords: ITS, recommender systems

Using intelligent tutoring system (ITS) data from industry partners, we aim to devise a question recommender model that will efficiently utilize student study time. That is, we want to only recommend questions that will improve a student's understanding of a concept, giving questions that are neither too simple nor too hard. This study highlights previous research on ITS recommender models and gives an analysis of industry data. Since most ITSs are proprietary, we hope to illuminate what data is likely available to these recommender systems and what kind of variation is seen within the data. Therefore, among our analysis we consider different measurements of question difficulty and compare to expert classification as well as determine if time taken to answer a question is a reliable measurement and should be used in our model.

I10. Topic Modeling on Bach Chorales
Yingsi Jian, Undergraduate, Department of Statistics
Co-Authors: Rayleigh Lei, Long Nguyen
Keywords: topic modeling, music, and projected data
Topic modeling is an unsupervised machine learning technique for discovering latent topics across text documents. Each topic represents a distribution over words, and each document is a mixture of topics. We are interested in applying topic modeling on musical data because these patterned structures may also be present in musical pieces. To perform topic modeling on musical data, we propose several notions of words to represent different musical elements. We apply these ideas and recent advances in topic modeling to J.S. Bach's chorales and discuss the different themes discovered.

Oral Presentations I

O1a. Methods to Account for Uncertainty in Latent Class Assignments when using Latent Classes as Predictors in Regression Models, with Application to Acculturation Strategy Measures

Zhangchen Zhao, PhD Candidate, Department of Biostatistics
Co-Authors: Michael R. Elliott, Bhramar Mukerjee, Alka Kanaya, Belinda L. Needham
Keywords: Latent Class Analysis, Measurement Error, Bayesian Models, Depression

Latent class models have become a popular means of summarizing survey questionnaires and other large sets of categorical variables. Often these classes are of primary interest to better understand complex patterns in data. Increasingly, these latent classes are reified into predictors of other outcomes of interests, treating the most likely class as the true class to which an individual belongs even there is uncertainty in class membership. This uncertainty can be viewed as a form of measurement error in predictors, leading to bias in the estimates of the regression parameters associated with the latent classes. Despite this fact, there is very limited literature treating latent class predictors as measurement error models. Most applications ignore this issue and fit a two-stage model that treats the modal class prediction as truth. Here we develop two approaches - one likelihood-based, the other Bayesian - to implement a joint model for latent class analysis and outcome prediction. We apply these methods to an analysis of how acculturation behaviors predict depression in South Asian immigrants to the US. A simulation study gives guidance for when a two-stage model can be safely implemented and when the joint model may be required.

O1b. Variational Inference of Joint Models using Multivariate Gaussian Convolution Processes

Xubo Yue, PhD Pre-candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Raed Al Kontar
Keywords: multivariate Gaussian convolution process, Cox model, variational inference, sparse approximation

We present a non-parametric prognostic framework for individualized event prediction based on joint modeling of both longitudinal and time-to-event data. Our approach exploits a multivariate Gaussian convolution process (MGCP) to model the evolution of longitudinal signals and a Cox model to map time-to-event data with longitudinal data modeled through the MGCP. Taking advantage of the unique structure imposed by convolved processes, we provide a variational inference framework to simultaneously estimate parameters in the joint MGCP-Cox model. This significantly reduces computational complexity and safeguards against model overfitting. Experiments on synthetic and real world data show that the proposed framework outperforms state-of-the art approaches built on two-stage inference and strong parametric assumptions.

O1c. Effects of statistical adjustments for subgroup differences in non-probability sample web surveys

Fernanda Alvarado-Leiton, PhD Pre-candidate, Program in Survey Methodology
Co-Authors: Sunghee Lee

Given the rapid growth of internet usage, web surveys have become a popular platform for survey data collection. Advantages of web data collection such as timeliness and relatively low costs make web surveys an attractive option for many researchers and survey organizations. However, an important drawback is that, more often than not, web surveys are implemented in non-probability, volunteer samples; thus, limiting their inferences to target populations. To overcome this obstacle, statistical adjustments are usually performed. However, it is not clear how effective these adjustments are for population subgroups. Particularly, when racial/ethnic disparities are of interest, potential biases in non-probability-based Web survey data and the degree to which statistical adjustments may address the biases in racial/ethnic disparities.

To fill this gap in the literature, this study examines health disparities related to race/ethnicity from a volunteer-sample Web survey with and without statistical adjustments and compares them to benchmarks, disparities estimates from an area-probability sample survey. We hypothesize that health disparities are attenuated in the web survey because, in general, respondents volunteering to participate may be more similar to each other regardless of the race and ethnicity and this similarity may be stronger than their racial dissimilarities. Specifically, we will use the National Health Interview Survey (NHIS) as our benchmark data set and data from The Study on Perceptions of Health, a non-probability Web survey conducted by the Institute for Social Research at the University of Michigan in 2017. The Web survey sample is comprised of roughly equal numbers of non-Hispanic Whites, non-Hispanic Blacks, Hispanics whose dominant language is English and Hispanics whose dominant language is English. We will focus on disparities in variables such as self-rated health, the number of chronic conditions, health risk behaviors (e.g., smoking, exercising) and health care usage. We will perform propensity score adjustments for the volunteer survey data based on a wide range of variables, including, demographics (e.g., age, gender, education, language), socio-economic (e.g., immigration, housing ownership) and health variables (e.g., disabilities). We then compare the results to the estimates from NHIS to examine if the racial/ethnic disparities are attenuated in the web volunteer sample and whether adjustments produce estimates closer to the benchmarks.

Oral Presentations II

O2a. Active Learning for Nonparametric Regression Using Purely Random Trees
Jack Goetz, PhD Candidate, Department of Statistics
Co-Authors: Ambuj Tewari, Paul Zimmerman
Keywords: Active Learning, Regression

Active learning is the task of using labelled data to select additional points to label, with the goal of fitting the most accurate model with a fixed budget of labelled points. In binary classification active learning is known to produce faster rates than passive learning for a broad range of settings. However in regression restrictive structure and tailored methods were previously needed to obtain theoretically superior performance. In this paper we propose an intuitive tree based active learning algorithm for non-parametric regression with provable improvement over random sampling. When implemented with Mondrian Trees our algorithm is tuning parameter free, consistent and minimax optimal for Lipschitz functions.

O2b. Functional Principal Component Analysis for Extrapolating Multi-stream Longitudinal Data
Seokhyun Chung, PhD Pre-candidate, Department of Industrial & Organizational Engineering
Co-Authors: Raed Al Kontar
Keywords: Functional Principal Component Analysis, Gaussian Process, Multi-stream, Longitudinal Data

The advance of modern sensor technologies enables collection of multi-stream longitudinal data where multiple signals from different units are collected in real-time. In this article, we present a non-parametric approach to predict the evolution of multi-stream longitudinal data for an in-service unit through borrowing strength from other historical units. Our approach first decomposes each stream into a linear combination of eigenfunctions and their corresponding functional principal component (FPC) scores. A Gaussian process prior for the FPC scores is then established based on a functional semi-metric that measures similarities between streams of historical units and the in-service unit. Finally, an empirical Bayesian updating strategy is derived to update the established prior using real-time stream data obtained from the in-service unit. Experiments on synthetic and real-world data show that the proposed framework outperforms state-of-the-art approaches and can effectively account for heterogeneity as well as achieve high predictive accuracy.

O2c. Scalable Information Estimation for Deep Neural Networks

Morteza Noshad, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Yu Zeng, Alfred Hero

Keywords: Mutual Information, Deep Learning, Deep Neural Networks, Non-parametric Estimation

The Mutual Information (MI) is an often used measure of dependency between two random variables utilized in information theory, statistics and machine learning. Recently several MI estimators have been proposed that can achieve parametric MSE convergence rate. However, most of the previously proposed estimators have high computational complexity of at least O(N^2).

We propose a unified method for empirical non-parametric estimation of general MI function between random vectors in R^d based on N i.i.d. samples. The reduced complexity MI estimator, called the ensemble dependency graph estimator (EDGE), combines randomized locality sensitive hashing (LSH), dependency graphs, and ensemble bias-reduction methods. We prove that EDGE achieves optimal computational complexity O(N), and can achieve the optimal parametric MSE rate of O(1/N) if the density is d times differentiable. To the best of our knowledge EDGE is the first non-parametric MI estimator that can achieve parametric MSE rates with linear time complexity. We illustrate the utility of EDGE for the analysis of the information plane (IP) in deep learning. Using EDGE we shed light on a controversy on whether or not the compression property of information bottleneck (IB) in fact holds for ReLu and other rectification functions in deep neural networks (DNN).

Oral Presentations III

O3a. Use of simulations to study the impact of linkage disequilibrium in the distribution of genetic risk scores across populations.
Pedro Orozco del Pino, PhD Pre-candidate, Department of Biostatistics
Co-Authors: Sebastian Zöllner
Keywords: Statistical genetics, genetic risk scores, simulation study

Genetic risk scores (GRS) quantifies an individual's risk of developing a disease based on his/her genetic information. Its applications range from statistical genetics methods to diagnosis and treatment of disease. However, most GRS have been built with information from European population samples. Therefore, may not have enough predictive power in other populations, which could result in healthcare inequities. Linkage Disequilibrium (LD) structure is one of several factors that may influence the differences between populations of GRS distribution. We present a simulation study that uses 1000 Genome Project haplotypes to quantify how LD structure can predict differences in GRS distributions between populations. We present an example that simulates GWAS studies in two populations with different LD structure and compares their GRS distribution under different scenarios. We show that even small differences in populationâ€™s genetics can substantially impact the GRS distribution.

O3b. Anomaly Detection in Partially Observed Traffic Networks
Elizabeth Hou, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Yasin Yilmaz, Alfred Hero
Keywords: Quantile Regression, High-Dimensional Statistics, Misspecification, Model-Free

This paper addresses the problem of detecting anomalous activity in traffic networks where the network is not directly observed. Given knowledge of what the node-to-node traffic in a network should be, any activity that differs significantly from this baseline would be considered anomalous. We propose a Bayesian hierarchical model for estimating the traffic rates and detecting anomalous changes in the network. The probabilistic nature of the model allows us to perform statistical goodness-of-fit tests to detect significant deviations from a baseline network. We show that due to the more defined structure of the hierarchical Bayesian model, such tests perform well even when the empirical models estimated by the EM algorithm are misspecified. We apply our model to both simulated and real datasets to demonstrate its superior performance over existing alternatives.

O3c. On the Optimality of Perturbations in Stochastic and Adversarial Multi-armed Bandit Problems

Baekjin Kim, PhD Candidate, Department of Statistics
Co-Authors: Ambuj Tewari
Keywords: Bandits, Follow-The-Perturbed-Leader, Perturbation, Thompson Sampling, sub-Weibull, Extreme value theory, Discrete choice theory

We investigate the optimality of perturbation based algorithms in the stochastic and adversarial multi-armed bandit problems. For the stochastic case, we provide a unified analysis for all sub-Weibull perturbations. The sub-Weibull family includes sub-Gaussian and sub-Exponential distributions. Our bounds are instance optimal for a range of the sub-Weibull parameter. For the adversarial setting, we prove rigorous barriers against two natural solution approaches using tools from discrete choice theory and extreme value theory. Our results suggest that the optimal perturbation, if it exists, will be of Frechet-type.

Poster Session I

P1a. Monitoring High-dimensional Data Streams
Zheng Gao, PhD Candidate, Department of Statistics

We discuss the problem of detecting and locating sparse signals in high dimensional data streams. Our message is two-fold.

First, we argue that cross-sectional dependence can be much more valuable than temporal dependence under memory constraints. Second, we explain why classical PCA-based methods remain competitive compared to variants based on robust PCA methods when signals are sparse.

We propose, and apply, a simple online method to the motivating application of distributed denial of service (DDoS) attacks detection on Internet traffic data streams. The method is implemented as a module in a Internet traffic monitoring system, running in real time for threat detection and identification.

P1b. Prediction for an individual's risk of PsA before symptoms appear using genetic data
Sunyi Chi, Master Student, Department of Biostatistics
Co-Authors: Matthew Patrick, Kevin He, Alex Tsoi
Keywords: Machine learning, Prediction, Psoriasis

Psoriatic arthritis (PsA) is a complex chronic musculoskeletal condition that occurs in ~30% of psoriasis patients. Current approaches to PsA diagnosis are based on clinical, laboratory and radiological features. There is limited systematic strategy to provide quantitative assessment for PsA risk among psoriasis patients, before symptoms appear and utilizes the differences in genetic architecture between PsA and cutaneous-only psoriasis (PsC) to assess PsA risk before symptoms appear. Here, we introduce a computational pipeline combining statistical and machine learning method for predicting PsA among psoriasis patients using genetic differences between psoriasis subtypes to assess risk of PsA and improve time efficiency to make it affordable to use in practice.

P1c. Semi-Supervised Sequence Learning using Deep Generative Models with Applications to Healthcare Data

Weijing Tang, PhD Candidate, Department of Statistics
Co-Authors: Ji Zhu
Keywords: Semi-supervised learning, Recurrent neural network, Generative model

Deep neural network classification models have been increasingly used to analyze large-scale electronic health records data and shown superior prediction performances. In general, the success of these models relies on the accessibility of a large number of labeled training data. In many healthcare settings, however, only a small number of accurately labeled data is available while unlabeled data is abundant. Further, input variables such as laboratory tests and charted events in the medical setting are usually sequential or longitudinal in nature, which poses additional challenges. In this project we propose new semi-supervised sequence learning methods, using deep generative models, to leverage both labeled and unlabeled data. We apply these methods to 5 mortality-related binary classification problems on a benchmark dataset extracted from the public MIMIC III database, and demonstrate that the proposed semi-supervised learning methods outperform supervised methods that use labeled data only.

P1d. Asymptotic Independent U-Statistics in High-Dimensional Adaptive Testing

Yinqiu He, PhD Candidate, Department of Statistics
Co-Authors: Gongjun Xu, Chong Wu, Wei Pan

Many high dimensional hypothesis testings examine the moments of the distributions that are of interest, such as testing of mean vectors and covariance matrices. We propose a framework that constructs a family of U statistics as unbiased estimators of those moments. In this talk, the usage of the framework is illustrated by testing for independence. We show that under null hypothesis, when both data dimension and sample size go to infinity, U statistics of different finite orders are asymptotically independent and normally distributed. Moreover, they are also asymptotically independent of the max-type test statistic, whose limiting distribution is an extreme value distribution. Based on the asymptotic independence property, we construct an adaptive testing procedure which combines p values computed from U statistics of different orders. Since higher order U statistics are usually more powerful against sparse alternatives and lower order U statistics are usually more powerful against dense alternatives, this adaptive procedure is powerful against different alternatives.

P1e. The Effect of Mutation Subtypes on the Allele Frequency Spectrum and Population Genetics Inference

Kevin Liao, Master Student, Department of Biostatistics
Co-Authors: Jedidiah Carlson, Sebastian Zoellner
Keywords: Population genetics, Allele frequency spectrum, Selection, Demographic inference, Mutation subtypes

The allele frequency spectrum (AFS) is a summary of genetic variation in a population that is commonly used for population genetics inference such as testing for selection and inferring demographic history. However, mutational mechanisms such as biased gene conversion and mutation rate heterogeneity are known to operate on specific sequences. As a result, the AFS can differ drastically depending on nucleotide context. Currently, no systematic review exists of how nucleotide context affects the AFS and downstream population genetics inference.

From whole genome sequencing data, we constructed 96 mutation subtypes (MST) from the adjacent nucleotides of a point mutation and constructed their AFS. Summary statistics of the AFS and various parameter estimates were then inferred and compared across MSTs. For each subtype, we quantified biased gene conversion and estimated population genetics parameters using DaDi. We reaffirm that higher mutation rates artificially lower the true singleton count and biased gene conversion leads to a systematic increase in intermediate allele frequencies. Furthermore, AFS-based demographic inference is strongly influenced by local nucleotide context. Estimates of the population genetics parameter theta varied by almost two orders of magnitude and exponential growth rate differed drastically amongst subtypes.

Current AFS-based inference may be biased due to averaging the AFS of all mutation types in a genomic region to construct the frequency spectrum. For example, tests of selection in local regions can have an increased rate of false positives and loss of power by failing to consider the nucleotide composition of the window. Future work will attempt to modify popular population genetic inference methods to account for local nucleotide context.

P1f. A Hierarchical Bayesian Approach to Neutron Spectrum Unfolding with Organic Scintillators

Haonan Zhu, PhD Pre-candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Yoann Altmann, Angela Di Fulvio, Stephen McLaughlin, Sara Pozzi, Alfred Hero
Keywords: Organic scintillators, Spectral unfolding, Bayesian Inference, Markov-chain Monte Carlo methods

We propose a hierarchical Bayesian model and state-of-art Monte Carlo sampling method to solve the unfolding problem, i.e., to estimate the spectrum of an unknown neutron source from the data detected by an organic scintillator. Inferring neutron spectra is important for several applications, including nonproliferation and nuclear security, as it allows the discrimination of fission sources in special nuclear material (SNM) from other types of neutron sources based on the differences of the emitted neutron spectra. Organic scintillators interact with neutrons mostly via elastic scattering on hydrogen nuclei and therefore partially retain neutron energy information. Consequently, the neutron spectrum can thus be derived through deconvolution of the measured light output spectrum and the response functions of the scintillator to monoenergetic neutrons. The proposed approach is compared to three existing optimization-based methods using simulated data to enable controlled benchmarks. We consider three sets of detector responses. One set corresponds to a 2.5 MeV monoenergetic neutron source and two sets are associated with (energy-wise) continuous neutron sources (252Cf and 241AmBe). Our results show that the proposed method has similar or better unfolding performance compared to existing approaches in terms of accuracy and robustness against limited detection events, while requiring less user supervision. The proposed method also provides a posteriori confidence measures, which offers additional information regarding the uncertainty of the measurements and the extracted information.

P1g. Efficient estimation of multi-dimensional linear discriminator

Debarghya Mukherjee, PhD Candidate, Department of Statistics
Co-Authors: Ya'acov Ritov, Moulinath Banerjee
Keywords: Non-parametric, High-dimension, Non-standard problems

Manski's celebrated maximum score estimator for the censored response linear model has been the focus of much investigation in both economics and statistics literature, but its behavior in the growing dimension still remains largely unknown. This project seeks to address the gap. Two different cases are considered: p grows with n but at a slower rate (i.e. p/n -> 0) and p>>n (fast growth). By relating Manski's score estimation to empirical risk minimization in a classification problem, we studied the convergence properties under suitable margin condition. We have also established minimax bounds under both regimes, which differs by a log factor. In slow growth regime, we have constructed an estimator which is minimax optimal. Finally, we provide some computational recipes for the maximum score estimator in growing dimensions that shows promising results.

P1h. Sample Size Considerations for Comparing Dynamic Treatment Regimens in a SMART with a Repeated-Measures Outcome

Nick Seewald, PhD Candidate, Department of Statistics
Co-Authors: Daniel Almirall
Keywords: Dynamic treatment regimen, Sequential multiple assignment randomized trial, Randomized trial, Longitudinal data, Sample size

Clinicians and researchers are increasingly interested in how best to individualize interventions. A dynamic treatment regimen (DTR) is a sequence of pre-specified decision rules which guide the delivery of a course of treatments that is tailored to the changing needs of the individual. The sequential multiple-assignment randomized trial (SMART) is a research tool that can be used to inform the construction of effective DTRs. We introduce sample size formulae for SMARTs in which the primary aim is to compare two embedded DTRs using a continuous repeated-measures outcome collected at three time points throughout the study. The method is based on a longitudinal analysis that accounts for unique features of a SMART, including modeling constraints and the over/under-representation of different sequences of treatment among participants. We also extend the method to choose both sample size and the number of measurement occasions in a SMART in order to maximize statistical power subject to a budget constraint. We illustrate the method using ENGAGE, a SMART aimed at developing a DTR for re-engaging patients with alcohol and/or cocaine use disorders who have dropped out of treatment.

P1i. Climate Change and Impacts on Air Pollution -- Results from North China

Ziping Xu, PhD Pre-candidate, Department of Statistics
Co-Authors: Song Xi Chen, Xiaoqing Wu
Keywords: Climate change, PM2.5

There are speculations that the severe air pollution experienced in North China were the acts of climate change in general and a decreasing northerly wind in particular. We first conduct a retrospective analysis on 38 years (1979-2016) reanalyzed meteorological data from ERA-Interim, an archive of European Centre for Medium-Range Weather Forecasts (ECMWF) to quantify meteorological changes over the 38 years. Statistically significant changes have been detected in the surface temperature, relative humidity and boundary layer height in the region between the first and the second 19-year periods from 1979 to 2016. However, there was no significant reduction in the northerly wind within the mixing layer. We then build regression models of PM2.5 on the meteorological variables using the 2015 and 2016 observations at 32 cities of the study region, which are used to quantify effects of the meteorological changes between the two 19-years periods on PM2.5. It is found that the average meteorological changes led to 2% to 7% reduction in monthly PM2.5 averages in most cities.

P1j. Testing complex multivariate mediation hypotheses
Joseph Dickens, PhD Candidate, Department of Statistics

Mediation analysis is used to test whether an intermediate mechanism either partially or fully explains an association between an exposure and outcome measure. In practice, a researcher could have several exposures and outcomes as well as multiple potential mechanisms explaining the associations. We develop a methodology for studying mediating relationships between vector-valued exposures, mediators and outcomes. We evaluate the structural hypotheses using likelihood ratio tests, using constrained maximum likelihood estimation to estimate model parameters under the null hypothesis. The null parameter space for many of the structural hypotheses is a smooth manifold that intersects itself. Where the manifold intersects itself, the sampling distribution of the likelihood ratio test statistic is non-trivial. We apply existing theory for testing hypotheses at singularities to characterize the sampling distribution of the likelihood ratio test statistic. We numerically demonstrate that, for reasonable sample sizes and problem dimension, our likelihood ratio tests both control their level and have reasonable power. Our work broadens the scope of structural questions a practitioner can ask and answer in the context of mediating variables.

P1k. Bayesian Dose-Finding Designs using Pharmacokinetics (PK) for Phase I Clinical Trials
David Todd, Masters, Department of Biostatistics

A phase I dose-finding study is usually the first trial in human subjects and has the goals of evaluating the safety of a tested pharmaceutical drug. This often offers the first analysis in describing the dose-concentration and dose-toxicity response relationships.
However, even if dose-finding and pharmacokinetic analysis are carried out in the same trial, they are often independently analyzed, due to the increased complexity of simultaneously analyzing both Thus it is worth determining what additional information on the dose-toxicity relationship is gained by incorporating pharmacokinetic information in the dose allocation process, and whether the benefit is sufficient to balance the added complexity of such a trial. To help simplify this process, a R-package called 'dfpk' was published in 2018 that implements novel methods for dose-finding phase I clinical trials incorporating pharmacokinetic data in the dose-toxicity relationships. An extensive summary of the 'dfpk' R-package, including its main functions, and how it can help improve the design and implementation of phase I clinical trials is examined.

P1l. Multilevel Regression and Poststratification with Unknown Population Distributions of Poststratifiers

Katherine Li, Masters, Department of Biostatistics
Co-Authors: Yajuan Si
Keywords: Multilevel Regression and Poststratification, Bootstrap, Imputation, Predictive

Multilevel regression and poststratification (MRP) can stabilize small area estimation via hierarchical model smoothing and adjust for the sample non-representativeness by post-stratifying to the population information. However, the population distribution of the poststratifiers may not be available. We propose flexible, nonparametric methods to impute a poststratifier's values for unsampled units and integrate the imputation uncertainty with the finite population inference of interest under a systematic Bayesian framework. We use simulation studies to demonstrate the bias and efficiency gains of imputing the unsampled poststratifiers under MRP comparing with alternative approaches.

P1m. Nonresponse in Online Panel

Wenshan Yu, Masters, Department of Survey Methodology

Sample surveys are supposed to support finite population inference. However, in the past 30 years, decreasing response rate has become a nonignorable threat for this function. Moreover, a more significant concern is that whether respondents and non-respondents differ in some characteristics, such that the difference between the two population can bias the estimates. Given its importance, the nonresponse mechanism has long been a critical area in survey research.

I aim to utilize data from the Understanding America Study to help explain the nonresponse mechanism. My research question for the proposed study is what the selectivity is in the panelists of a probability-based online panel (Understanding America Study) within the following two stages 1) agreement to initially participate in the panel, 2) wave-to-wave participation.

Poster Session II

P2a. A Longitudinal Analysis of Politically Active Twitter Users
Robyn Ferg, PhD Candidate, Department of Statistics
Co-Authors: Johann Gagnon-Bartsch, Fred Conrad

Relationships found between data extracted from social media and public opinion polls have led to optimism about supplementing traditional surveys with new sources of data. We provide evidence that a political signal is present in Twitter data. To do this, we follow politically active Twitter users over time. After classifying politically active users as a Democrat or a Republican, we demonstrate that there is a political signal in both tweeting frequency and sentiment of the politically active users, with a clear change in sentiment immediately following the 2016 presidential election. We follow these users through mid-2018 and find relationships between the sentiment of these users' tweets and presidential approval.

P2b. Automated Model Selection within Sequential Imputation of Missing Data for High-Dimensional Data Sets
Micha Fischer, PhD Candidate, Department of Survey Methodology
Keywords: Multiple Imputation, Missing Data, High-Dimensional Data Sets, Non-Normal Variables

Multiple imputation with sequential regression models is often used to impute missing values in data sets and leads to unbiased results if the missing data is missing at random and models are correctly specified. However, in data sets where many variables are affected by missing values, proper specifications of those sequential regression models can be burdensome and time consuming, as a separate model needs to be developed by an human imputer for each variable. Even available software packages for automated imputation procedures (e.g. MICE, IVEware) need model specifications for each variable containing missing values. Additionally, their default models can lead to bias in imputed values, for example when variables are non-normally distributed.

This research aims to automate the process of sequential imputation of missing values in high-dimensional data sets consisting of potentially non-normally distributed variables. The proposed algorithm performs model specification by selecting the best imputation model from several parametric and non-parametric models in each step of the sequential imputation procedure. The best imputation model for an outcome variable achieves the highest similarity between imputed and observed values after conditioning on the response propensity score for the outcome variable. A simulation study investigates in which situations this automated procedure can outperform the usual approaches (MICE, IVEware, human imputer). Preliminary results will be presented here.

P2c. Power and tuning for the knockoff filter
Brook Luers, PhD Candidate, Department of Statistics
Keywords: Knockoff filter, Variable selection, False discovery rate, Collinearity

The knockoff filter is a variable selection technique for linear regression with finite-sample control of the regression false discovery rate (FDR). The regression FDR is the expected proportion of selected variables which, in fact, have no effect in the regression model. To control the regression FDR, the knockoff filter constructs synthetic variables which mimic the observed covariates but are known to be irrelevant to the regression. Constructing these synthetic variables involves tuning parameters which can increase collinearity and reduce power. In this poster, I describe conditions under which the knockoff filter amplifies collinearity and propose alternative criteria for selecting tuning parameters in order to limit collinearity. Simulation results indicate that these new tuning choices do not reduce the power of the knockoff filter when using the lasso and can improve power in ridge and ordinary least squares regressions.

P2d. A Bayesian Mixture Model to Estimate the Effect of an Ordinal Predictor
Emily Roberts, PhD Pre-candidate, Department of Biostatistics
Co-Authors: Lili Zhao
Keywords: Bayesian method, Ordinal predictors

In the medical field, ordinal predictors are commonly seen in regression analysis. Often, ad-hoc approaches are used to analyze these variables. For example, many treat these predictors as categorical by ignoring ordering, as continuous by assuming that the ordered values are equally spaced, or as dichotomous based some threshold for convenient decision making. We propose a Bayesian mixture model to automate such decisions. In situations where a true threshold exists, the method is able to determine the optimal cutoff value for the predictor. Since the model can simultaneously assess the appropriate form of the predictor and perform estimation, arbitrary thresholds and multiple testing concerns are avoided. By using a mixture model, the outcome is estimated as a weighted average of linear and thresholding functions of the predictor. This method is applicable to continuous, binary, and time-to-event outcomes, readily extends to multiple predictors with different numbers of ordered levels, and adjusts for confounding variables. We demonstrate the model's accuracy through simulation studies and apply it to real datasets with binary and survival outcomes.

P2e. Precision Matrix Estimation with Noisy and Missing Data
Roger Fan, PhD Candidate, Department of Statistics
Co-Authors: Byoungwook Jang, Yuekai Sun, Shuheng Zhou
Keywords: High-dimensional, Optimization, Covariance estimation, Graphical models

The estimation of the dependency graphs of graphical models is one of the most relevant problems in modern statistics, and when data are fully observed penalized methods like the graphical Lasso and node-wise regressions have been standard for this estimation under sparsity conditions. There are extensions of these methods to more complex data with noise, dependence, and/or missing values; however, in these settings, the relative performance of different methods is not well understood and algorithmic gaps still exist. In particular, Loh and Wainwright (2015) show how the graphical Lasso can be modified to handle missing and noisy data. In high-dimensional settings, however, this requires solving the graphical Lasso objective with a indefinite input matrix, presenting novel optimization challenges. We develop an alternating direction method of multipliers (ADMM) algorithm for this problem, providing a feasible algorithm to estimate precision matrices with indefinite input and potentially nonconvex penalties. We compare this method with existing alternative solutions and empirically characterize the tradeoffs between them. Finally, we illustrate the usage of this method by exploring the networks between US senators estimated from voting records data.

P2f. Comparison of Generalized Estimating Equations and Random Forest in Estimating Causal Effects
Hengshi Yu, PhD Pre-candidate, Department of Biostatistics
Keywords: Causal effect, Generalized estimating equation, Random forest

In longitudinal data analysis with treatment assignment, the treatment effects are often constant with respect to the treatment values. Sometimes, however, the causal effects might also depend on the treatment value and are not linear for the expected potential outcome to the treatment value. In this project, we used generalized estimating equations (GEE) in longitudinal data analysis as well as random forest (RF) in statistical learning to estimate the treatment-varying treatment effect. We constructed estimators and compared them in simulation studies. The original GEE method shows its well performance in estimating the treatment effects over the state-of-the-art methods in statistical learning.

P2g. Supervised Principal Component Analysis via Manifold Optimization
Alexander Ritchie, PhD Candidate, Department of Electrical Engineering & Computer Science
Co-Authors: Clayton Scott, Laura Balzano
Keywords: Dimensionality Reduction, Manifold Optimization, Supervised Learning

High dimensional prediction problems are pervasive in the scientific community. In practice, dimensionality reduction (DR) is often performed as an initial step to improve prediction accuracy and interpretability. Principal component analysis (PCA) has been utilized extensively for DR, but does not take advantage of outcome variables inherent in the prediction task. Existing approaches for supervised PCA (SPCA) either take a multi-stage approach, lack a direct means of trading off variation explained and prediction accuracy, or both. We present a manifold optimization approach to SPCA that simultaneously solves the prediction and dimension reduction problems, in both the regression and classification settings. Our empirical results show that the proposed approach explains nearly as much variation as PCA while outperforming existing methods in prediction accuracy.

P2h. Neighborhood Effects on Health Outcomes: Findings from Mexico City
Tahmeed Tureen, Masters, Department of Biostatistics
Co-Authors: EFS Roberts, MM Tellez-Rojo, BN Sanchez
Keywords: Built Infrastructure, Physical Activity, Neighborhoods, Mexico City

A growing body of research has investigated the associations between built infrastructure and physical activity in high-income countries (HIC) like the United States (US). The research has shown that higher availability and quality of infrastructure is associated with higher physical activity among adults. Research conducted among adolescents is less conclusive in general, and little work in this area exists in Latin America (LA). Most countries in LA are low-to-middle income (LMIC), and have infrastructures that are markedly different from the US due to economic differences but also due to different histories of urban settlements and community building. Using data from a longitudinal cohort and objective measures of built infrastructure, this project investigates the association between features of the infrastructure and physical activity among adolescents in working class neighborhoods in Mexico City, Mexico (CDMX). Features of the built infrastructure investigated are: availability of sidewalks, green cover (trees on streets), public transportation, and informal business activity on streets (mobile vendors). Results from this research bring to question whether associations found in HIC are transportable to LMIC. Given the expected population growth and rapid rate of urbanization in LA, it is pressing to identify features of built infrastructure that support and promote health-related behaviors in LA.

P2i. Applications of Propensity Score Methods for Several Outcomes of Interest Using Claims Data

Ryan Ross, Masters, Department of Biostatistics
Co-Authors: Megan Caram, Paul Lin, Min Zhang, Bhramar Mukherjee
Keywords: Claims data, Causal inference, Propensity score, Prostate cancer

Electronic Health Record (EHR) and medical and insurance claims are now common data sources to answer a variety of questions in biomedical research. While extensive, these datasets are observational, which limits effective understanding of interventions and differences between groups being compared. Several methods have been developed to better estimate causal treatment effects, often utilizing the propensity score. This paper offers a comprehensive guide to researchers in using propensity methods for estimating causal treatment effects on several types of outcomes common to medical studies, such as time to event and time varying outcomes. The methods are illustrated using a cohort of patients with prostate cancer from the Clinformatics TM Data Mart Database (OptumInsight, Eden Prairie, Minnesota).

P2j. Water as an economic good: does water price reflect water scarcity?

Xiaodan Zhou, Masters, Department of Statistics

Water resources provide many ecological services for human beings. It is not only essential to our every lives and agricultural and industrial activities, but also to critical ecological functions. As a renewable yet limited resource, water availability is challenged by a growing population and increasing consumption. Therefore, water scarcity ranks among the most critical challenges in the 21st century. As a solution to water scarcity, water pricing helps conserve and allocate water, therefore promoting the water use efficiency and sustainability. Many studies analyzed how water use responds to different water pricing schemes in different sectors. However, few of them connect water price and water scarcity directly to investigate (1) whether the long-term water scarcity is reflected by present water price, and (2) how the water scarcity information is incorporated in the water pricing process. The paper aims at filling this gap. Here we analyzed the correlation between local water scarcity and water price in 10 U.S. states at the county level. The regression results reveal that there is no significant correlation between them. A detailed economical and political explanatory analysis is followed to explain the phenomena.

P2k. A Bayesian Method for High-dimensional Discrete Graphical models

Anwesha Bhattacharyya, PhD Candidate, Department of Statistics
Co-Authors: Yves Atchade

This work introduces a Bayesian methodology for fitting large discrete graphical models with spike-and-slab priors to encode sparsity. We consider a quasi-likelihood approach that enables node-wise parallel computation resulting in reduced computational complexity. We introduce a scalable langevin MCMC algorithm for sampling from the quasi-posterior distribution which enables variable selection and estimation simultaneously. We present extensive simulation results to demonstrate scalability and accuracy of the method. We also analyze the 16PF dataset to illustrate performance of the method.

Notes