2016

Michigan Student Symposium for Interdisciplinary Statistical Sciences

Program of Events

March 25, 2016
8:30 AM – 7:00 PM
Rackham Building

MSSISS 2016 Student Organizing Committee:

Wenting Cheng (Biostatistics)

Kristjan Greenewald (Electrical Engineering &Computer Science)

Wenbo Sun (Industrial & Operations Engineering)

Joon Ha Park and Naveen Naidu Narisetty (Statistics)

Felicitas Mittereder (Survey Methodology)

Faculty Advisory Committee:

Dr. Timothy D. Johnson (Biostatistics)

Dr. Clayton Scott (Electrical Engineering &Computer Science)

Dr. Eunshin Byon (Industrial & Operations Engineering)

Dr. Susan Murphy (Statistics)

Dr. Brady West (Survey Methodology)

Sponsors:
Biostatistics, Electrical Engineering & Computer Science, Industrial & Operations Engineering, Statistics and Survey Methodology departments at the University of Michigan

We thank Dr. Susan Murphy and Dr. Brady West for their many thoughtful suggestions. We thank Lorie Kochanek for her help in organizing this event, Jessica Darga for her help in constructing MSSISS website and Mei Mei for her help in designing MSSISS logo.

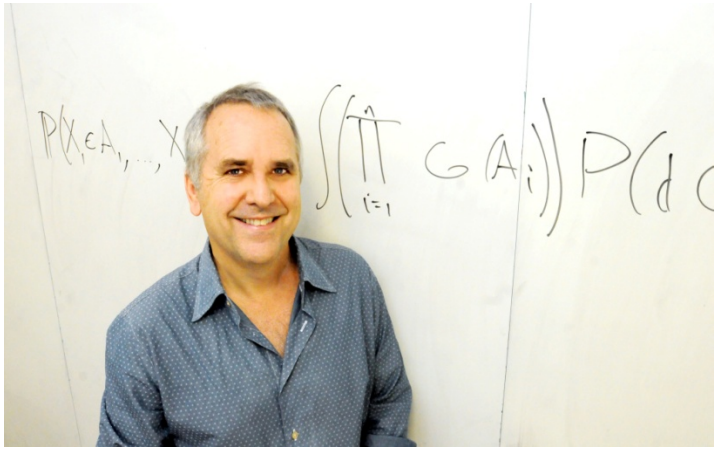
The MSSISS 2016 Student Organizing Committee also acknowledges Michigan Institute for Data Science (MIDAS) for its generous support.

MIDAS MICHIGAN INSTITUTE FOR DATA SCIENCE UNIVERSITY OF MICHIGAN

Schedule

8:30 – 8:55	Registration and Breakfast (Assembly Hall)
8:55 – 9:00	Welcoming Remarks (Amphitheater)
9:00 – 10:15	Student Oral Presentations I (Amphitheater) Colleen A. McClain, Department of Survey Methodology Mathieu Bray, Department of Biostatistics Pin-Yu Chen, Department of EECS
10:15 – 11:15	Poster Session I (East Conference Room)
11:15 – 12:30	Student Oral Presentations II (Amphitheater) Jesus Daniel Arroyo, Department of Statistics Karen Nielsen, Department of Statistics Michael Hornstein, Department of Statistics
12:30 – 1:45	Poster Session II and Lunch (Assembly Hall)
1:45 – 3:00	Student Oral Presentations III (Amphitheater) Sayantan Das, Department of Biostatistics Fan Wu, Department of Biostatistics Mingdi You, Department of IOE
3:00 – 4:00	Poster Session III and Coffee Break (West Conference Room)
4:00 – 4:55	Keynote Address: Professor Michael I. Jordan, University of California, Berkeley (Amphitheater) *On the Computation/Statistics Interface*
4:55 – 5:00	Closing Remarks and Student Awards (Amphitheater)
5:00 – 7:00	Reception sponsored by MIDAS (Assembly Hall)

Keynote Speaker



Dr. Michael I. Jordan

Dr. Michael I. Jordan is the Pehong Chen Distinguished Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics at the University of California, Berkeley. His research in recent years has focused on Bayesian nonparametric analysis, probabilistic graphical models, spectral methods, kernel machines and applications to problems in signal processing, statistical genetics, computational biology, information retrieval and natural language processing. Prof. Jordan was elected a member of the National Academy of Sciences (NAS) in 2010, of the National Academy of Engineering (NAE) in 2010, and of the American Academy of Arts and Sciences in 2011. He is a Fellow of the American Association for the Advancement of Science (AAAS). He has been named a Neyman Lecturer and a Medallion Lecturer by Institute of Mathematical Statistics (IMS). He is a Fellow of the IMS, a Fellow of the IEEE, a Fellow of the AAAI, and a Fellow of the ASA.

On the Computation/Statistics Interface

There are many research challenges and open problems at the interface of computation and statistics. In this talk I'll discuss some recent progress at this interface, including (1) the use of concurrency control mechanisms in the setting of statistical inference; (2) an analysis of the computational complexity of high-dimensional Bayesian variable selection; and (3) a new variational perspective on the mysterious phenomenon of Nesterov acceleration

Data Challenge Competition

In addition to the usual research presentations, MSSISS 2016 student committee is pleased to announce a data challenge event in a partnership with the Michigan Data Science Team (MDST) and the Michigan Institute for Data Science (MIDAS).

This year's data challenge featured a particularly interesting dataset. The MDST has compiled records of every fatal car accident reported in the United States between 2003 and 2014, a dataset known as the Fatal Accident Reporting System Dataset, or FARS. The challenge was to predict whether or not a drunk driver was involved in the accident. Cash prizes totaling $1,000 are awarded at MSSISS 2016 to the teams achieving the best performance and/or devising the most innovative statistical methods.

The winning team will present their method on a poster in the third poster session in the West Conference Room.

Student Oral Presentations

Session I Amphitheater 9:00 AM – 10:15 AM

Pathways to consent: Psychosocial characteristics, biomarker collection, and data linkage in the Health & Retirement Study

Colleen A. McClain, Department of Survey Methodology
Co-authors: Sunghee Lee

Increasing collection of auxiliary information within the survey context—including consent for physical measurements, biomarker collection, and permission to link to administrative records—has led to substantial gains in researchers' ability to draw conclusions from a wealth of data points. However, the need to consent respondents separately for these complex, possibly intrusive requests has the potential to introduce nonresponse bias at each stage of the consent process. A relatively new body of research has assessed the correlates of consent to such auxiliary requests (e.g. Sakshaug, Couper, & Ofstedal, 2010; Sakshaug, Couper, Ofstedal, & Weir, 2012; McClain, Lee, & Faul, 2014) as well as methods to increase consent rates (e.g. Sala, Knies, & Burton, 2014; Kreuter, Sakshaug, & Tourangeau, 2015). Fewer studies have investigated the *mechanisms* behind such complex decisions.
We use data from the 2010 Health and Retirement Study to investigate the role that one respondent attribute—psychosocial orientation—plays in the consent process. In doing so, we first identify a set of traits that may relate to such decisions, building on potential psychological foundations of survey response summarized by Groves (1989). Importantly, we expand upon the limited set of Big Five personality measures that the few prior studies in this domain have focused on, zeroing in on particular traits we hypothesize as related to consent and to our health and economic outcomes of interest. Preliminary results, including those from structural equation models incorporating relevant demographic and health-related correlates, display the differential role of psychosocial orientation for each consent request examined. We address implications for understanding consent mechanisms and improving methods of nonresponse adjustment, and discuss future directions for research incorporating the interviewer-respondent interaction as well as testing specific hypotheses about consent-related biases in resulting data.

Incorporating Candidates with Multiple Associated Incompatible Donors in Kidney Paired-Donation

Mathieu Bray, Department of Biostatistics
Co-authors: Wen Wang, Peter X-K Song, John D. Kalbfleisch

Kidney paired-donation (KPD) represents one avenue for transplant candidates seeking a donor. Candidates with a willing incompatible donor are matched with other pairs in an effort to find combinations of donor exchanges that allow all candidates involved to obtain transplants. In some cases, candidates have several donors, all incompatible, willing to participate in KPD. Having multiple donors in KPD introduces additional opportunities to match with other pairs, and also allows for fallbacks to immediate alternatives should failures (eg. withdrawal, laboratory test overturning presumed compatibility) occur in a determined transplant

arrangement. Exchanges involving pairs with multiple donors should be preferred amongst possible arrangements in KPD. We formulate an objective assignment of expected utility for KPD exchange combinations, both with and without recourse to available fallback options and accounting for probabilities of failure, where candidates can have multiple associated incompatible donors. This extends previous mathematical formulations of KPD (Li, 2012). Further, we illustrate through simulation the benefits for candidates in seeking out multiple donors for KPD.

AMOS: A Model Order Selection Criterion for Spectral Graph Clustering

Pin-Yu Chen, Department of EECS
Co-authors: Alfred Hero

One of the longstanding open problems in unsupervised classification is the so-called model order selection problem: automated selection of the correct number of classes or clusters. In the context of spectral graph clustering (SGC), this is equivalent to the problem of finding the number of connected components or communities in an undirected graph. We propose a solution to the SGC model selection problem under a homogeneous random interconnection model (RIM) using a novel selection criterion that is based on an asymptotic phase transition analysis. Our solution can more generally be applied to discovering hidden block diagonal structure in symmetric non-negative matrices. Numerical experiments on simulated graphs validate the phase transition analysis, and real-world network data is used to validate the performance of the proposed model selection procedure

Student Oral Presentations

Session II Amphitheater 11:15 AM – 12:30 PM

Community detection in networks via Sparse Principal Components Analysis
Jesus Daniel Arroyo, Department of Statistics
Advisor: Elizaveta Levina

Community detection in networks, the problem of finding groups of nodes that have more connections to each other than to the rest of the network, has received a lot of attention in the literature, but many methods only allow for a node to belong to exactly one community. In practice, nodes in network may belong to multiple communities, or to no communities at all. Here we propose a new efficient algorithm for overlapping community detection based on sparse principal component analysis. The algorithm has a computational cost similar to that of estimating the largest eigenvectors of the adjacency matrix, and does not require an additional clustering step like spectral clustering techniques. We show that our method is consistent in selecting the community memberships under an overlapping version of the stochastic blockmodel and evaluate the method empirically on simulated and real-world networks, showing good statistical performance and computational efficiency.

Capitalizing on the Use of Basis Sets in Regression Spline Mixed Models
Karen Nielsen, Department of Statistics
Advisor: Rich Gonzalez

Disciplines have their favorite or conventional basis set. For example, polynomials are common in psychology, and Fourier transforms are often used in engineering and physics. There exists a vast set of possibilities for basis sets in generalized splines, which can be used to impose expected structure on a model via piecewise sums of any functions. The properties of basis set transformations can be leveraged in powerful ways, especially when they can give model parameters more natural, domain-relevant interpretations. Here, we will show how Regression Spline Mixed Models (RSMM) can combine the nonparametric features of splines with a hierarchical random effects framework to explore EEG data at any of the many levels that are collected and of interest to researchers. We will then show how a verbalized hypothesis can be translated into a basis set for formal testing of interpretable model parameters. Having the ability to work at these levels in any time series biological context (EEG, fMRI, MEG, EKG, pupilometry, and others) is useful for recognizing outliers, learning about variance and statistical significance, and inspiring further analyses or future studies.

Efficient mean structure estimation using matrix variate data

Michael Hornstein, Department of Statistics
Co-authors: Kerby Shedden, Shuheng Zhou

Recent work has discovered previously unappreciated correlations between observational units in large biological data sets, leading to interest in matrix-variate data with row and column correlations. Consider matrix-variate data X, for which the covariance of the vectorized data, vec(X), can be decomposed into a Kronecker product of matrices A and B. Suppose the samples are divided into two groups, and we are interested in estimating the differences in group means for each variable. We present two methods based on generalized least squares for estimating the mean structure as well as covariance matrices in this setting. The first method is based on group centering each column of X before estimating the row wise covariance matrix B. The second method uses a model selection step which allows us to perform group centering only on genes with sufficiently large effect size, while leaving others to the usual (default) global centering procedure. We provide rates of convergence on mean and covariance estimation. Our analysis applies to a model where the marginals of each row and column vectors follow subgaussian distribution under the same mean structure.

Student Oral Presentations

Session III Amphitheater 1:45 PM – 3:00 PM

Next Generation Imputation Methods
Sayantan Das, Department of Biostatistics
Co-authors: Christian Fuchsberger, Goncalo R Abecasis

Genotype imputation is a key step in human genetic studies. Modern technologies have led to a rapid increase in the size of publicly available reference panels rendering contemporary imputation tools computationally cumbersome. In this presentation, I will describe novel statistical tools for faster genotype imputation that scale fairly with large reference panels. First, we take advantage of local haplotype redundancy to dynamically reduce the Hidden Markov Model state space. This leads to significant savings in physical memory and CPU time (100 fold), while preserving the same accuracy. Next, we apply a more aggressive version by collapsing haplotypes that are identical only at positions where the sample was genotyped, leading to further substantial savings. Our methods have been implemented in a freely available software (minimac3). We also developed a web imputation server that simplifies analyses for users by automating the process and eliminating the need for cumbersome data access agreements, thus allowing researchers to devote their time to more interesting tasks. Our web server, to date, has been used to process >2.4 million genotyped samples and has over 800 users.

A Pairwise Likelihood Augmented Estimator for Left-Truncated Data with Time-Dependent Covariates
Fan Wu, Department of Biostatistics
Co-authors: Sehee Kim, Yi Li

The study of the long-term survival of end stage renal disease patients using the transplant registry is challenging due to left-truncated survival time and time-dependent covariates of interest. Ignoring the information in the truncation times, conventional conditional approaches yield consistent estimates but are less efficient. We introduce a semi-parametric estimation method for the Cox model under left-truncation that shows substantially improved efficiency. Rather than impose parametric forms on the truncation distribution, we rely on a pairwise likelihood argument to eliminate it. The greatest advantage of this approach is that, even when the time-dependent covariate depends on the truncation time, the proposed estimator is still accurate with better precision. Large sample properties have been shown using techniques in empirical process and U-process, and finite sample properties based on the asymptotic normality with a closed-form variance estimator have been demonstrated by extensive simulation studies. The proposed method is applied to the OPTN/UNOS kidney transplant data.

Statistical Models for Characterizing Heterogeneous Wake Effects in Multi-turbine Wind Farms

Mingdi You, Department of IOE
Co-authors: Eunshin Byon, Judy Jin, Giwhyun Lee

Wind turbines in a wind farm exhibit heterogeneous power generations due to wake effects. Because upstream turbines absorb kinetic energy in wind, downstream turbines have power deficits. Moreover, the power deficit at downstream turbines shows heterogeneous patterns, depending on weather conditions. This study introduces a new approach for characterizing heterogeneous wake effects based on the Gaussian Markov random field (GMRF). A case study demonstrates the proposed approach's superior performance over commonly used alternative methods.

Poster Presentation

Session I East Conference Room 10:15 AM – 11:15 AM

Poster 1a.
Minimum Predictive Risk Subspace Selection in Misspecified Quantile Regressions
Alexander Giessing, Department of Statistics
Advisor: Xuming He

We introduce the concept of minimum predictive risk subspaces for variable selection in misspecified quantile regressions. This framework is natural in situations in which the true quantile functions are unknown or stipulating the existence of a single true quantile function is too restrictive. Our selection procedure adjusts the selection bias associated with the relative entropy of the candidate predictors. Under mild conditions on the quantile functions, the new procedure is shown to consistently select the set of predictors that span a low-dimensional linear subspace with minimal predictive risk. This property proves particularly effective when analyzing heteroscedastic and skewed data. Numerical studies demonstrate the advantages (as well as challenges) of the new method on the basis of various misspecified quantile regressions models.

Poster 1b.
Characterizing Gastrointestinal Activity in the Stomach and Small Intestine
Joseph Dickens, Department of Statistics
Co-authors: Kerby Shedden, Jason R. Baker, Allen Lee, Gordon Amidon, Duxin Sun, William L. Hasler

Migrating motor complexes (MMC) regulate gastrointestinal activity during fasted periods between meals. MMCs are subdivided into phases defined by frequency features and contraction strength and duration. Of particular interest is MMC phase III during which regular contractions move the intestinal contents toward the colon. We use data collected during intubation studies that record manometric data for up to twelve hours at up to nine locations in the stomach and the small intestine. Our preliminary results suggest that phase III events exhibit two types of distinct of behavior: (a) periods of coordinated, anterograde contractions and (b) periods of less organized, possibly retrograde, contractions. Propulsion velocities are estimated during periods of organized activity. Using functional principal component analysis, we describe phase III contraction shape and observe substantial variation in the timing of maximum contraction pressure. We also develop techniques for detecting phase III events and use our algorithm to produce estimates of the survival curve for waiting times between distinct phase III events.

Poster 1c.
Detecting the source of DNA contamination in genotyping arrays and correcting for it
Gregory JM Zajac, Department of Biostatistics
Co-authors: L. G. Fritsche, S. L. Dagenais, R. H. Lyons, C. M. Brummett, G. Abecasis

Genotyping arrays measure the relative intensities of allele specific probes to call the genotype at each locus. Contamination, the mixture of DNA samples prior to genotyping, increases the probability that the calling software fails to call a genotype, reducing the power of genetic analyses. Current methods for contamination detection do not find which samples contributed DNA. We propose a two-step process. First, we use a particular sample's alternate allele intensity and the genotypes of all potential contamination sources to identify DNA donors. Then, we remove the contamination from those sources to recover the original genotype. Experimental results from intentionally mixed HapMap samples show that our method estimated contamination more accurately than existing methods VerifyIDintensity and BAFRegress, and correction performs well at up to 10% contamination. Applying our method in a large genotyping study helped identify the step where contamination occurred. This approach has several advantages, including more accurate estimation of the contamination proportion. Detecting the source of contamination provides useful information to guide improvements in sample preparation.

Poster 1d.
Stochastic Optimization for high-dimensional Mixed Effect Generalized Linear Models
Jun Guo, Department of Statistics
Co-authors: Yves F. Atchade

Motivated by recent genome-wide association studies in admixed/trans-ethnic population, we are interested in estimation and inference of high dimensional mixed effect generalized linear models. These models are widely applicable to many fields, including gene mapping studies of quantitative, binary and categorical disease outcomes or phenotypes of interest in the presence of various confounding effects. The high dimensional models we concern often possess non-convex and non-smooth objective functions. We propose a new stochastic perturbed ADMM algorithm to estimate the model. Simulation studies show that our algorithm performs comparably to a new stochastic proximal gradient algorithm proposed by Atchade et al. 2014. We develop the non-convex and stochastic convergence theory for the proximal algorithm, which initializes our route for statistical and computational theory development of the proposed methods.

Poster 1e.
Deep Learning In Mobile Health
Michael Kovalcik, Department of EECS
Advisor: Susan Murphy

The primary purpose of this study was to investigate whether regression analysis done by a deep learning neural network is a good approach to estimating the value function in reinforcement learning. We are also evaluating how useful deep learning may be in a setting with smaller data sets such as the randomized trials in mobile health. Our evaluation will involve a neural network

with fewer nodes and hidden layers to accommodate the small data sets. We will consider the use of simulated trials to initialize our starting parameters as a way of dealing with less data. The simulated trial data means we will need less training data to for the algorithm to learn the correct model.

Poster 1f.
A Web-Based Treatment Assignment System for Research Trials
Yumeng Li, Sophie Yu-Pu Chen, Department of Biostatistics
Co-authors: Kerby Shedden, Brenda W Gillespie

"Minimization" is a common way to balance potential confounders when assigning subjects sequentially to treatments in a trial, but several practical challenges result from the need to randomize subjects in real time at their point of enrollment. Therefore, we have implemented a web-based minimization and trials management system that is open source, and is designed to be hosted on Google Appengine, a low cost "platform as service." Our system allows the user to specify the number treatment arms, the number of stratification variables, the number of levels within each variable, and the randomization ratios. Various forms of logging, data retention, and data export are supported, and the system can be configured to allow multiple users to perform assignments for each trial. Google accounts are used to securely restrict access to each project to its authorized personnel, as defined by the project owner. In several years of use our application has never approached the resource limit for free Appengine accounts, allowing us to run it without incurring charges from Google.

Poster 1g.
Additive covariance model
Seyoung Park, Department of Statistics
Co-Authors: Shuheng Zhou, Kerby Shedden

In this paper, we study additive covariance models to explain two-way dependence in data. The baseline Kronecker sum covariance structure has the form of $\Sigma = A \oplus B := A \otimes I_n + I_m \otimes B$, where $A \in R^{m \times m}$ and $B \in R^{n \times n}$ are positive definite matrices, and $I_n \in R^{n \times n}$ is an identity matrix. Here, the matrix A describes the covariance structure between the columns of data, and the other covariance matrix B describes the covariance structure between the rows of data. This structure naturally arises from the following two-way dependence model: $X = X_0 + W$, where the rows of X_0 are independent having covariance matrix A, and the columns of W are independent with covariance matrix B. With a replicate of the data $X' = X_0 + W'$, where W' is a replicate of W , more general structure assumptions such as two-way dependence can be applied to each random part. We propose to use nodewise regression approach and the projected graphical lasso (GLasso) to estimate the covariance matrices A and B and their inverses. We apply the method to a neuromotor control study of hawkmoths (Sponberg et al. 2015) and analyze the temporal and spatial dynamics.

Poster 1h.
Fast Approximation of Small p-values in Permutation Tests by Partitioning the Permutation Space
Brian Segal, Department of Biostatistics
Co-authours: Hui Jiang, Thomas Braun

Researchers in genetics and other life sciences sometimes use permutation tests to evaluate differences between groups. Permutation tests have desirable properties, including exactness, and are applicable even when the distribution of the test statistic is analytically intractable. However, permutation tests can be computationally intensive. We propose an algorithm for quickly approximating small permutation p-values in two-sample tests. Our approach is based on a stochastic ordering of test statistics across partitions of the permutation space, which allows us to calculate p-values in partitions that require less computation and then predict p-values in partitions that would require more computation. In this presentation, we describe our method and demonstrate its use through simulations and an application to cancer genomic data. We find that our method is faster than a current leading method, and can successfully identify up- and down-regulated genes. We have implemented our method in the R package fastPerm.

Poster 1i.
Beverage-Specific Binge Drinking Patterns in Young Adults Aged 19/20
Stephanie A. Stern, Department of Survey Methodology
Co-authors: Yvonne M. Terry-McElrath, Megan E. Patrick

Objective: The prevalence of alcohol use among young adults has been widely researched, including binge drinking prevalence rates. However, beverage-specific binge drinking prevalence rates for beer, wine coolers, liquor, and wine are not yet documented for this age group. This study aims to describe the substances consumed at binge levels by young adults aged 19/20. Method: Data from the national Monitoring the Future study, which starts with annual 12th grade cohorts, were collected one or two years after high school in 2004-2014 (n=2216). Logistic regression was used to examine associations between beverage-specific binge drinking (i.e., 5+ drinks of each specified beverage) and gender, race/ethnicity, parent education, college status, and cohort year. Results: Overall binge drinking in the past two weeks was reported by 30.9% of young adults. Beverage-specific binge drinking was most common with liquor (22.3%) and beer (21.9%), followed by wine (4.5%) and wine coolers (3.0%). Males were significantly more likely to binge drink overall than females, as were White young adults as compared to their Black and Hispanic peers. Overall binge drinking was significantly and positively associated with respondents reporting either being four-year college students or having college-educated parents. The likelihood of binge drinking overall decreased between the 2002-2005 and 2010-2013 cohort groups, albeit not significantly (p=0.05). Beverage-specific binge drinking also differed by demographics. Males were significantly more likely to binge on beer and liquor, whereas females were significantly more likely to binge drink wine coolers and wine. Both two-year college students and young adults who did not attend college were significantly more likely than four-year college students to consume wine coolers at binge levels. The 2002-2005 cohorts were significantly more likely to consume beer at binge levels than the 2010-2013 cohorts. Conclusions: Different beverage-specific binge drinking patterns were observed by gender as

well as by college enrollment. Additionally, the likelihood of drinking beer at binge levels was lower for recent cohorts. (Funded by NIAAA R01AA023504 & NIDA R01 DA016575)

Poster 1j.
Evaluating use of a Cox regression model in Landmark analysis to approximate an Illness-Death Model
Krithika Suresh, Department of Biostatistics
Co-authors: Jeremy Taylor, Alexander Tsodikov

The landmarking approach to dynamic prediction incorporates time-dependent information accrued during follow-up to improve survival prediction probabilities. For several chosen landmark times, a dataset is created containing patients still at risk for failure and their covariate values at that time. These datasets are stacked and a simple Cox model is fit to the residual time. To account for "overlap" between the datasets, the Cox model coefficients are restricted to have a parametric form that is a function of the landmark time. Our goal is to determine whether a Cox model is appropriate to achieve consistency between the predictions at the different landmark times. Considering an illness-death model, we wish to find the form of the corresponding residual time model in a landmarking approach and assess whether it is consistent with a Cox model. By equating the residual time distributions under both approaches we identify the structure of the Cox model baseline hazard and covariate effects that corresponds to the landmark illness-death model. In the landmark setting, the covariate effects should be independent of the residual time. Since a simple Cox regression does not satisfy this condition, we explore alternative flexible model structures that provide a consistent and valid approximation.

Poster 1k.
Nonparametric Group Sequential Methods for Detecting Short-Term Survival Benefit from Multiple Follow-up Windows
Meng Xia, Department of Biostatistics
Co-Authors: Nabihah Tayob, Susan Murray

Interim statistical analyses have a long and respected history in clinical trial design for both ethical and financial reasons. Under similar train of consideration, health economists have emphasized for many years that patients placed more value on the short-term outlook of survival projections. Following this thought, in this research we developed group sequential methods for monitoring a clinical trial via Tayob and Murray's statistic which evaluated the behavior of 1-year restricted mean estimated from multiple, overlapping 1-year follow-up windows. Asymptotic joint distribution of Murray and Tayob test statistics calculated at different interim analysis times was studied for later on defining stopping boundaries based on the nature of the trial and desired operating characteristics. We will then focus on studying asymmetric bounds which are useful when consequences of stopping early are different according to the treatment difference that is emerging and hence avoid the ethically uncomfortable scenario of trial termination only after statistical proof of increased mortality on the experimental treatment arm.

Poster 1l.
Functional Data Analysis on Business Cycle Indicators
Jia Xiang, Department of Statistics
Advisor: Edward Rothman

Currently many investors and investing institutions are utilizing business cycle indicators in investment decision-making. I started with examining the effectiveness of current leading and coincident business cycle indicators in terms of the timing of the turning points and relative direction of movements with regard to the real GDP. I developed algorithms in R to apply smoothing functions, time warping, and continuous registration to collected data. I discovered that the Industrial Production Index is a desirable coincident indicator while S&P 500 and GDP are switching roles in leading and are also subject to Stock Volatility and Consumer Confidence.Take one step further than FDA, I proposed the ratio of the Industrial Production Index to the real GDP as a novel and flexible business cycle predictor.

Poster 1m.
Generating Correlated Synthetic Binary Indicators of Radio Listening Behavior for Long-Term Projections
H.Yanna Yan, Survey Methodology
Co-authors: Michael Elliott, Brady T. West, William Waldron

Radio panels, such as the Nielsen Audio panel, are a reliable source of information on the population's radio listening behavior. Nielsen panels consist of a representative probability sample of a target population, and panel members constantly carry a wearable device that extracts radio signals 24 hours a day. This device generates a binary variable (0/1) to indicate if a respondent listened to a particular radio station for a given 15-minute time window. This raises the question of can limited capture of exposure data (e.g., one week) be used to project radio listening behavior for longer periods of time? This study evaluates a proposed beta-binomial modeling approach for generating correlated synthetic indicators of radio listening behavior for an entire month based on one week of panel data. The presentation compares synthetic estimates of "cumulative reach" (i.e., number of people were exposed to a radio station) for one month to direct estimates of reach based on a full month of panel data, and finds that the proposed methodology works quite well.

Poster 1n.
Robustness of the Contextual Bandit Algorithm to Learning Effects
Xige Zhang, Department of Statistics
Co-authors: Huitain Lei

The contextual bandit algorithms have been widely used in many fields including online advertising and news recommendation. They have been recently proposed to form Just-In-Time Adaptive Intervention (JITAI) in mobile health. One of the most fundamental assumptions in contextual bandit problem is that the distribution of contexts is independent of actions at previous decision points. This assumption is fragile in real world mobile health problems, for there are many ways that the distributions of contexts change dynamically and are impacted by

past treatments. In this project, we investigate the performance of actor critic contextual bandit algorithm when learning effect presents. In mobile health, learning effects are one common form of how previous actions influence the distribution of future context. For example, people can form positive habits after receiving interventions on mobile phones. Our simulation study reveals that performance of the actor critic contextual bandit algorithm does not deteriorate significantly in the presence of learning effect.

Poster 1o.
Logistic regression model estimation and prediction incorporating coefficients information
Wenting Cheng, Department of Biostatistics
Co-authors: Jeremy M.G. Taylor, Bhramar Mukherjee

We consider a situation where there is a rich amount of historical data available for the coefficients and their standard errors in a logistic regression model $\text{logit}(\Pr(Y = 1|X)) = X\beta$ from large studies, and we would like to utilize this summary information for improving inference in an expanded model of interest, $\text{logit}(\Pr(Y = 1|X, B)) = (X, B)\Upsilon$, in a new dataset of moderate size. By using logistic regression approximation proposed by Monahan and Stefanski, 1992, we formulate the problem into an inferential framework where the historical information is translated in terms of a set of non-linear constraints on the parameter space. We propose several frequentist and Bayes solutions. For Bayes solutions, these non-linear constraints are treated as informative priors for β and weakly informative Cauchy priors for Υ (Gelman et al, 2008). We show that the transformation approach proposed in Gunn and Dunson, 2005 is a simple and effective computational method to conduct Bayesian inference in this situation. Our simulation results comparing these solutions indicate that historical information on model $\text{logit}(\Pr(Y = 1|X)) = X\beta$ can boost the efficiency of estimation and enhance prediction ability in the model of interest $\text{logit}(\Pr(Y = 1|X, B)) = (X, B)\Upsilon$.

Poster Presentation

Session II Assembly Hall 12:30 PM – 1:45 PM

Poster 2a.
Comparison of predictive accuracy among joint models of longitudinal and survival data
Guanming Zheng, Department of Statistics
Advisor: Walter Dempsey,Susan A. Murphy

Survival studies often generate not only a survival time for each patient but a sequence of health measurements while the patient remains alive. These outcomes are often separately analyzed; however, often a joint modeling approach is required to yield deep understanding of the underlying mechanisms. Various methods have been proposed for fitting joint models for longitudinal and time-to-event data; here we compare two – the shared random-effects model and the joint latent-class model – on a variety of survival studies in order to assess each in terms of predictive accuracy. The main goal is a better understanding of if and when the model classes outperform each other in the task of prediction.

Poster 2b.
Statistical Methods for Rare Variant Test with Multiple Phenotypes
Diptavo Dutta, Department of Biostatistics
Co-authors: Seunggeun Lee

In genetic association studies, joint testing of related phenotypes can provide novel insights into the genetic architecture of complex diseases. Although several methods exist for multi-phenotype tests with common variants, only a few exist for rare variants. To address this, we present several strategies to combine multi-phenotypes into gene-based tests, specifically a PC-based approach, regression-model-based approach, and omnibus test approaches that combine the two. The PC-based approach modifies the SKAT-O test based on the principal components of the phenotypes and aggregates signals for multiple PCs. The regression-based approach models the effect-sizes of the variants through correlations in mixed models and conducts a variance component test. From extensive simulation studies, we show that these tests can improve power over standard single-phenotype tests, while maintaining type1 error. Their relative performance depends on the number of associated phenotypes and correlation patterns. The omnibus test generally has robust power regardless of the genetic model. We applied our methods to data on nine amino-acid phenotypes from METSIM studies to identify associated variants.

Poster 2c.
Detecting differentially expressed metabolic pathways with adjustments for macronutrient intake
Teal Guidici, Department of Statistics
Co-authors: George Michailidis

Differential expression testing and set enrichment analysis are commonly used to summarize the results of high throughput biological experiments, to generate biologically meaningful hypothesis for further analysis. and to aid in the planning of validation experiments. Conventional approaches to differential expression testing and set enrichment analysis do not usually account for individual variation in relevant background features, in many cases due to lack of pertinent data. These features are especially relevant in the context of metabolomics, where blood metabolite levels can react sensitively and quickly to changes in nutrient intake. In this project we introduce a network based method for detecting differentially expressed metabolites and metabolic pathways, while adjusting for individual variation in the consumption of relevant macronutrients through the integration of nutrition intake data. We test our method on metabolomic and nutrition intake data from a controlled feeding study featuring two distinct diets (a high polyunsaturated fat diet and a high carbohydrate diet).

Poster 2d.
Dynamical Communication for Adaptive mHealth Interventions
Kelly Hall, Department of Statistics
Co-authors: Shawna Smith, Gaurav Paruthi, Jin Seok Andy Lee, Susan Murphy

Tailored communication delivered via mobile devices, using messages specific to demographic, psychosocial, and dynamic context variables, has proven effective for health behavior change. This tailoring framework suggests that the more appropriate messages are to a person and his or her context, the fewer barriers there will be to "buying in" to a treatment program. We aimed to create a library of context-based (location, time of day, day of week, and weather) health messages for a mobile application, HeartSteps, designed to decrease sedentary behavior – ultimately for patients transitioning out of cardiac rehabilitation programs. To expedite this processes, we used the crowdsourcing platform Amazon Mechanical Turk to generate and curate messages, creating a database of over 500. These messages were used during the six-week, forty-participant HeartSteps pilot study which concluded in February of 2016. Our talk will cover advantages and drawbacks of this crowdsourcing strategy for mHealth, participant response to various types of messages, and the potential for future, more context- and behavior-specific tailoring.

Poster 2e.
Sampling Methods to Improve the Efficiency of Two-phase Estimation
Paul Imbriano, Department of Biostatistics
Co-authors: Trivellore Raghunathan

A two-phase survey design is used when a variable, Y, is expensive or difficult to obtain, relative to variables auxilary variables X. Y could be a disease or biomarker. In the first phase variables

X's, are measured on a random sample of the population. In the second phase, the more costly Y is measured on a subsample of participants from phase I. Utilizing the phase I variables from all participants improves the efficiency in estimating Y over using only data from phase II participants. A regression estimator can be used in estimating the population mean of Y, or more generally, we can perform multiple imputation to estimate the mean, moments, and conditional relationships of Y. We propose new methods for selecting our phase II sample based on the values of X from phase I, that uses the distribution of the variance of the regression estimator and the distribution of the fraction of missing information. These selection methods improve the efficiency of the regression estimator and multiple imputation over using simple random sampling.

Poster 2f.
Network cross-validation by edge sampling
Tianxi Li, Department of Statistics
Co-authors: Elizaveta Levina, Ji Zhu

Many models and methods are now available for network analysis, but model selection and tuning remain challenging. Cross-validation is a useful general tool for these tasks in many settings, but is not directly applicable to networks since splitting network nodes into groups requires deleting edges and destroys some of the network structure. We propose a new edge sampling cross-validation strategy applicable to a wide range of network problems. We provide an error bound on cross-validated estimates in a general setting, and in particular show that the method has good asymptotic properties when selecting the number of communities under the stochastic block model. Numerical results on both simulated and real networks show that our approach performs well for a number of model selection and tuning parameter tasks.

Poster 2g.
Minimizing sum of truncated convex functions and its applications
Tzu-Ying Liu, Department of Biostatistics
Co-authors: Hui Jiang

We study a class of problems where sum of truncated convex functions is minimized. In statistics it is often encountered when L0-penalized models are fitted. While in general they often leads to NP-hard non-convex optimization problems, we show that there is a polynomial-time algorithm in low-dimensional settings by partitioning the domain of the function. Our algorithm shows superior performance when compared with other global optimization algorithms, especially in cases where the objective function has a complex landscape. We also demonstrate the utility of our algorithm for outlier detection in robust linear regression, and we find that it outperforms state-of-the-art methods when a large amount of outliers are present.

Poster 2h.
Predictive Models in Horticulture: A Case Study with Royal Gala Apples
Tom Logan, Department of IOE
Co-authors: S. McLeod, S. Guikema

Decision makers in horticulture want to forecast their crop characteristics. Predictions of the crop inform decisions which influence pricing, marketing, logistics, and even consumer satisfaction. This article summarises predictive horticultural models in the literature, and finds confusion exists between predictive and explanatory models. It encourages the use of statistical learning and nonlinear methods for future predictive models. Then it demonstrates how predictive models can be constructed using data for Royal Gala apples from orchards within New Zealand. For the eight years of data available the model has been shown to have a mean predictive error of 6.7%. The best model was an ensemble of a linear model, a BART, and a Boosted CART. Statistical learning techniques present substantial opportunity to the horticultural industry and to future attempts to develop more accurate predictive models.

Poster 2i.
Identification of gene pairs as biomarkers using penalized regression
Lan Luo, Department of Biostatistics
Co-authors: Hui Jiang

In recent years, gene pairs have been proposed as biomarkers for clinical studies. Instead of selecting individual genes as biomarkers, using gene pairs allows us to use the ratio of two genes and avoid the issues associated with data normalization. To select gene pairs, we fit regression models based on the log-transformed gene expression levels with an L1 penalty and an additional equality constraint that requires the sum of the coefficients to be zero. We use an Alternating Direction Method of Multipliers (ADMM) algorithm to fit the model. Inexact-ADMM is used to accelerate the optimization when possible. We also developed an R package that solves general equality-constrained L1-regularized general linear models via ADMM and inexact-ADMM.

Poster 2j.
Cluster-level dynamic treatment regimens and sequential, multiple assignment, randomized trials: Estimation and sample size considerations
Timothy NeCamp, Department of Statistics
Co-authors: Daniel Almirall, Amy Kilbourne

Individual-level dynamic treatment regimens (DTRs), also known as adaptive interventions, are used to sequence treatment decisions based on a person's changing health information. We introduce cluster-level DTRs where treatment decisions are made at the cluster level. We present cluster-level sequential, multiple-assignment, randomized trial (SMARTs) designs, designs used to develop high-quality DTRs, in which randomization is at the cluster level and outcomes are at the individual-level. We develop a weighted least squares regression approach to compare treatment regimes embedded in a cluster-randomized SMART. We also develop a sample size calculator for designing cluster-randomized SMARTs. The validity and robustness of the calculator under various settings, including when working assumptions are violated, are

evaluated through simulation. To illustrate our methods we utilize Adaptive Implementation of Effective Programs Trial (ADEPT), a cluster randomized SMART currently being conducted.

Poster 2k.
Random Intercept Bayesian Additive Regression Trees: Application to Longitudinal Prediction
Yaoyuan Vincent Tan, Department of Biostatistics
Co-authors: Carol A.C. Flannagan, Michael R. Elliott

Modeling non-linear associations between outcome and covariates including interactions, has long been areas of research in statistics. Classification and regression tree (CART) models are particularly convenient for modeling high-dimensional interactions; however, they are unable to handle non-linear relationships between outcome and covariates. Bayesian Additive Regression Trees (BART) maintain power to estimate high-dimensional interactions while addressing non-linearity by approximating the non-linear relationship using a sum of regression trees. The BART model was developed under the framework of independent subjects and cannot handle repeated measurements on the same subject. To account for repeated measurements, we extend BART by introducing a random intercept into the model. We provide two alternative distribution assumptions on the random intercept: normal and non-central t. We compared results from these two assumptions and found that random intercept BART (riBART) with a non-central t random intercept worked better for continuous outcomes while riBART with a normal random intercept worked better for binary outcomes. We then compared our better performing models with BART on two simulated repeated measurements dataset, one with continuous outcome and the other with binary outcome. We found that riBART performed better than BART in terms of mean squared error and area under the receiver operating characteristic curve for continuous and binary outcomes respectively. We also considered an application of riBART to predicting driver behavior when executing a left turn at intersections using a distance series of vehicle speeds.

Poster Presentation

Session III West Conference Room 3:00 PM – 4:00 PM

Poster 3a.
Exploring the Robustness of Risk-Adjusted Measures of Hospital Quality
Blake Arnold, Department of Statistics
Advisor: Edward Rothman

The purpose of this project is to explore patient risk factors for colectomy surgery and provide measures of hospital-level surgical quality, adjusted for these risk factors. In particular, surgical site infections are used to measure adverse patient outcomes from colectomy cases. Logistic regression, propensity score matching, and hierarchical mixed models are used to create measures of hospital quality after adjusting for preoperative patient risk factors. By using multiple methods to identify risky hospitals we determine the robustness of these estimates. This project will propose methods for hospital groups and governmental agencies to identify healthcare facilities where there are significant opportunities for surgical or operational improvement. These methods can be similarly applied to identify risk factors and underperforming hospitals for other surgical procedures in the future.

Poster 3b.
Singularity structures and parameter estimation in mixtures of skew normal distributions
Nhat Ho, Department of Statistics
Co-authors: XuanLong Nguyen

Finite mixures of skew normal distributions have become increasingly popular in recent years due to their flexibility in modeling asymmetric data. However, these models appear to suffer from various orders of singularities of mixing measure under both the exact-fitted and over-fitted setting. These singularities happen not only in the vicinity of symmetry but also in the setting of homologous sets, a new phenomenon due to the complex interaction among the parameters of the mixing measure. Apart from these singularities, skew normal mixtures also suffer from the influence of non-linear partial differential equation. It leads to two new ways of characterizing the singularity level of mixing measure. One way is based on the solvability of non-linear or inhomogeneous system of polynomial equations while the another way is based on the borrowing strength phenomenon among multiple systems of polynomial equations. The rich spectrum of the singularity structure of mixing measure consequently leads to various intricate degrees of parameter estimation under skew normal mixtures. Due to these inference issues, we propose a simple yet efficient adaptive surgery of these singularities to recover the optimal convergence rate n^{-1/2} of the parameter estimation. Finally, we carry out the careful simulation studies to illustrate all the results being developed in this talk.

Poster 3c.
Latent Laplacian Maximum Entropy Discrimination for Detection of High-Utility Anomalies
Elizabeth Hou, Department of Statistics
Co-authors: Alfred O. Hero, Kumar Sricharan

Data-driven anomaly detection methods suffer from the drawback of detecting all instances that are statistically rare, irrespective of whether the detected instances have real-world significance or not. In this paper, we are interested in the problem of specifically detecting anomalous instances that are known to have high realworld utility, while ignoring the low-utility statistically anomalous instances. To this end, we propose a novel method called Latent Laplacian Maximum Entropy Discrimination (Lat- LapMED). This method uses the EM algorithm to simultaneously incorporate the Geometric Entropy Minimization principle for identifying statistical anomalies, and the Maximum Entropy Discrimination principle to incorporate utility labels, in order to detect high-utility anomalies. We apply our method to both simulated and real datasets and demonstrate that our method has superior performance over existing alternatives including semi-supervised classification methods that attempt to detect the high-utility points without incorporating information about their statistical rarity, and unsupervised anomaly detection algorithms that simply detect statistically rare instances without incorporating utility labels.

Poster 3d.
Comparison and validation of statistical methods for predicting the failure probability of trees
Elnaz Kabir, Department of IOE
Co-Authors: Seth Guikema

This paper examines disparate statistical methods for predicting the failure likelihood of trees in the face of storms and also comparing their accuracies. Being able to make accurate predictions plays a key role in helping arborists to do preventive measures with the aim of decreasing the chance of failure or even cutting down the hazard trees. The data used consists of four factor variables including the location of each tree, the tree species, whether the tree was pruned and whether there are any removed trees around the tree, and also two continuous variables including diameter at breast height (DBH) and height. Different data mining methods are used to predict the failure probability of trees. They include logistic regression, random forest regression, classification and regression trees (CART), multivariate adaptive regression splines (MARS), artificial neural network (ANN), naïve-Bayes regression and an ensemble model. These models are validated through one hundred holdouts and the best ones in terms of accuracy are chosen for further analysis. Our results indicate that logistic regression, random forest and ensemble model of these two models predict the failure rate better than others.

Poster 3e.
Modeling multiple brain networks through linear mixed effects models
Yura Kim, Department of Statistics
Co-authors: Elizaveta Levina

Data on the brain's structural or functional connections is frequently represented in the form of networks, with a different network for each subject in the study. However, these networks all share the same set of nodes and can thus be analyzed jointly. Current work tends to either reduce them to global summaries such as modularity, or vectorize the edge values and ignore network structure. Here we propose a method for modeling brain networks via linear mixed effects models which takes advantage of the community structure, or regions, known to be present in the brain. The model allows us to compare different populations (for example, healthy and mentally ill patients) both globally and at the edge level, and find significant areas of difference. Further, we can incorporate the correlation between edges in the network inherent in brain data by allowing for a general variance structure in the mixed effects model. We illustrate the method by analyzing data from a study comparing schizophrenics to healthy controls.

Poster 3f.
Intelligent Sampling for Identifying Threshold in Observed Data-Bases and Time Series
Zhiyuan Lu, Department of Statistics
Co-authors: Moulinath Banerjee, George Michailidis

In threshold and location estimation one often tries to minimize over a set of location and auxiliary parameters, with the former adding computational time to what can be otherwise simple optimization schemes. We introduce a general 2-stage sampling scheme for observed databases and time-series, which we coin 'intelligent sampling', which aims to estimate parameters using an appropriately chosen subsample of the data. Ideally this method can yield optimal convergence rates using a vanishing fraction of the entire dataset, and our analysis of several change point problems support this hypothesis. Several single change-point scenarios with fixed signals are demonstrated where intelligent sampling achieves the optimal rate of convergence by only analyzing a subsample from the full dataset. Additionally, the possibility of adapting this methodology to nonparametric change point models, multiple change point models, change-planes models in higher dimensions, as well as thresholds of other kinds are also explored.

Poster 3g.
Assessing the Effect of Momentary Motivational Messages on Physical Activity in a Micro-Randomized Trial
Nicholas J. Seewald, Department of Statistics
Co-authors: Shawna N. Smith, Predrag Klasnja, Susan A. Murphy

Heart disease is the leading cause of death in the United States, yet many of these deaths can be prevented. Changes in lifestyle, including increased physical activity, are recommended by the Centers for Disease Control as a method of risk-reduction. In individuals who have recently experienced a cardiac event, however, rates of maintenance of such changes are quite low after exiting cardiac rehabilitation. To combat this, there is interest in developing individualized mobile health interventions delivered through a smartphone. The micro-randomized trial is an experimental framework which allows for the construction of effective “just-in-time” adaptive

interventions by repeatedly randomizing participants hundreds or thousands of times during the study. HeartSteps is a micro-randomized trial in which participants were randomly delivered motivational messages designed to encourage physical activity up to five times per day. Participant step counts were monitored using Jawbone Move fitness trackers. Here, we present preliminary results from HeartSteps which address the main, proximal effect on step count of delivering a suggestion versus not.

Poster 3h.
Learning network dynamics via regularized tensor decomposition
Yun-Jhong Wu, Department of Statistics
Advisor: Elizaveta Levina and Ji Zhu

Real networks often evolve over time, and interactions between nodes in networks are usually observed only at certain specific time points. In this work, we consider network data with time-stamped links. We propose to model such a dynamic network using a low rank tensor representation. This model characterizes time trends of multiple rank-1 factors and can be used to approximate more complicate networks. We develop an approach to fit this model based on a tensor completion algorithm and a smoothness penalty in the time domain, implemented with a highly salable power-iteration-based algorithm which can fit large sparse dynamic networks. The numerical experiments on simulated data as well as the Enron e-mail dataset demonstrate the potential of tensor methods for dynamic network data analysis.

Poster 3i.
Geometric Dirichlet Means Algorithm for Topic Inference
Mikhail Yurochkin, Department of Statistics
Co-Authors: XuanLong Nguyen

We propose a geometric algorithm for topic learning and inference that is built on the convex geometry of topics arising from the Latent Dirichlet Allocation model. To this end we study the optimization of a geometric loss function, which is a surrogate to the LDA's likelihood. Our algorithm makes use of k-means clustering as a building block, which overcomes the computational and statistical inefficiencies encountered by other techniques based on Gibbs sampling and variation inference, while achieving the accuracy comparable to that of a Gibbs sampler. The topic estimates produced by our method are shown to be statistically consistent. The algorithm is evaluated by extensive experiments on simulated and real data.

Poster 3j.
Assessing Factors Affecting Prices of Chinese Paintings with Particular Attention to Indirect Financial Market Movements
Lu Zhang, Department of Statistics
Advisor: Edward Rothman

I examine the market for five leading Chinese artists' paintings sold at auctions from 2008 to 2014. Using ordinary least squares regression supplemented with estimated derivatives using functional data analysis, I assess factors that play a significant role in the determination of selling price. In addition to traditional factors that are commonly used in the literature on art auctions

such as dimension, provenance, etc., I also studied features based on financial. There is well-documented evidence that the art market is related to financial markets, but little is known about the extent to which financial market performance and directional movements affect the price of a painting. To address this, financial markets' returns and their relative movements (predictors obtained from functional data analysis) are also included in the regression. Most hypotheses about the significance of different predictors, including indirect movements of financial markets, are validated.

Poster 3k.
Diffusion Model for Advanced Driver Assistance Systems Market Penetration Prediction
Shuning Zhang, Department of Statistics
Co-authors: Qi Luo

Diffusion model is a well-studied growth model used for innovations and technology forecasting. It consists of a simple differential equation describing the adoption of a new product entering the market. Advanced Driver Assistance System (ADAS) have been developed and implemented as a new automobile technology in recent years. We are interested in predicting its market share in the next decade with very limited information by assuming that it follows the diffusion model. This ADAS market penetration is important data support for an intelligent parking project in Mobility Transformation Center (known as Mcity). Compared to the airbag equipment data during 1990-1998, which is a representative passive safety feature in vehicles, it is observed that the data follows a typical diffusion model. ADAS belongs to active safety technology, which we believe to have similar patterns in growth. We fitted several growth models including Logistic Function, Gompertz Function and Bass Diffusion Model on the airbag historical data to select the best-fitted one. In the next step, we applied a text search tool on secondary cars market in order to investigate parameters in the diffusion model depending on different innovations such as cruise control and navigations. The two most significant parameters are the growth rate K and equilibrium height L. K is bounded by a regression model between K and average equipment cost P. L is scaled by the current fleet size comparing to the fleet size of 1998. With these results, we can estimate the percentage of cars in use which equip ADAS system in the next decade.

Poster 3l.
The Comparison of ACI and MCB Methods for Choosing a Set that Contains the Optimal Dynamic Treatment Regime
Rong Zhou, Department of Statistics
Co-authors: Tianshuang Wu

Dynamic treatment regimes (DTRs) have been used by clinicians for treatment decision-making at different stages for patient care. Sequential multiple assignment randomized trials (SMART) is a design to obtain data in order to find the optimal DTR that maximizes the expected cumulative outcome. Many methods have been developed to seek for the best DTR, including Q-learning methods and A-learning methods. In addition, a method called "Multiple Comparisons with the Best" is proposed by researchers to identify a set of DTRs that includes the optimal one. In this project, we focus on two methods: the modified ACI (Adaptive Confidence Intervals) method by Laber et al. and MCB (Multiple Comparisons with the Best) method by Ertefaie et al. We apply simulation in four different scenarios for both methods and report the results. By comparing the

probabilities that the best DTR is included into the constructed set, and the average set size of each method in four different scenarios, we conclude that we will recommend the MCB method by Ertefaie et al. in general.

Poster 3m.
Penalized Spline of Propensity Methods for Treatment Comparison
Tingting Zhou, Department of Biostatistics
Co-authors: Michael Elliott and Roderick Little

Observational studies lack randomized treatment assignment, and as a result valid inference about causal effects can only be drawn by controlling for confounders. When time dependent confounders are present that serve as mediators of treatment effects and affect future treatment assignment, standard regression methods for controlling for confounders fail. Similar issues also arise in trials with sequential randomization, when randomization at later time points is based on intermediate outcomes from earlier randomized assignments. We propose a Bayesian approach to causal inference in this setting called Penalized Spline of Propensity Methods for Treatment Comparison (PENCOMP), which builds on the Penalized Spline of Propensity Prediction method for missing data problems. The latter relies on the balancing property of propensity score to achieve double robustness by modeling the relationship between propensity scores and outcomes as a penalized spline regression. PENCOMP estimates causal effects by imputing missing potential outcomes with suitable spline models, and drawing inference based on imputed and observed outcomes. We demonstrate that PENCOMP has a double robustness property for causal effects, and simulations suggest that it tends to outperform doubly-robust marginal structural modeling when the relationship between propensity score and outcome is nonlinear or when the weights are highly variable.

Poster 3n.
Multiple Imputation of Missing Covariates for the Cox Proportional Hazards Cure Model
Lauren Beesley, Department of Biostatistics
Co-authors: Jonathan W Bartlett, Jeremy M G Taylor

We explore several approaches for imputing partially observed covariates when the outcome of interest is a censored event time and when there is an underlying subset of the population that will never experience the event of interest. We call these subjects “cured,” and we consider the case where the data are modeled using a Cox proportional hazards (CPH) mixture cure model. We study covariate imputation approaches using fully conditional specification (FCS). We derive the exact conditional distribution and suggest a sampling scheme for imputing partially observed covariates in the CPH cure model setting. We also propose several approximations to the exact distribution that are simpler and more convenient to use for imputation. A simulation study demonstrates that the proposed imputation approaches outperform existing imputation approaches for survival data without a cured fraction in terms of bias in estimating CPH cure model parameters. We apply our multiple imputation techniques to a study of patients with head and neck cancer.

Poster 3o.
Dynamic Metric Tracking
Kristjan Greenewald, Department of EECS
Co-authors: Stephen Kelley, Alfred Hero

Recent work in distance metric learning focused on learning transformations of data that best align with provided sets of pairwise similarity and dissimilarity constraints. The learned transformations lead to improved retrieval, classification, and clustering algorithms due to the more accurate distance or similarity measures. Here, we introduce the problem of learning these transformations when the underlying constraint generation process is dynamic. These dynamics can be due to changes in either the ground-truth labels used to generate constraints or changes to the feature subspaces in which the class structure is apparent. We propose and evaluate an adaptive, online algorithm for learning and tracking metrics as they change over time. We demonstrate the proposed algorithm on both real and synthetic data sets and show significant performance improvements relative to previously proposed batch and online distance metric learning algorithms.

Poster 3p.
Data Challenge Winners