MSSISS 2014

Michigan Student Symposium for Interdisciplinary Statistical Sciences

March 21, 2014
8:30am – 5:00pm
Rackham Building, 4th floor

Morning Session

8:30 – 8:55	Registration and Breakfast
8:55 – 9:00	Welcoming Remarks
9:00 – 10:15	Huitian Lei, Statistics *“A statistical decision procedure for personalizing treatment”*
	Xu Shu, Biostatistics *“Semiparametric methods to contrast restricted mean gap times”*
	Sandipan Roy, Statistics *“Estimating a change-point in high-dimensional Ising model”*
10:15 – 11:15	Poster Session I
11:15 – 12:30	Yang Liu, Electrical Engineering and Computer Science *“Anatomy of network-level malicious activities: Connectedness and inter-dependence”*
	Zhuqing Liu, Biostatistics *“Pre-surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model”*
	Pramita Bagchi, Statistics *“Inference for monotone trend under dependence”*

Afternoon Session

12:30 – 1:45	Poster Session II and Lunch
1:45 – 3:00	James Henderson, Statistics *“Network reconstruction using nonparametric additive ODE models”*
	Raphael Nishimura, Survey Methodology *“Item nonresponse in auxiliary variables used in calibration for survey sample data”*
	Laura Fernandes, Biostatistics *“Adaptive phase I clinical trial design using Markov models for conditional probability of toxicity in oncology”*
3:00 – 4:00	Poster Session III and Coffee Break
4:00 – 4:55	Keynote Address: Professor Jeff Wu, Georgia Institute of Technology *“From real world problems to esoteric research: Some examples”*
4:55 – 5:00	Closing Remarks and Student Awards

All of the events will take place on the fourth floor of the Rackham Building: oral presentations at the Rackham Amphitheatre*; breakfast, lunch, and coffee break at the Rackham Assembly Hall; poster sessions at East/West Conference Rooms* and Rackham Assembly Hall.

* No food is allowed in the Rackham Amphitheatre and East/West Conference Rooms.

Sponsors: Department of Biostatistics, Statistics, and Electrical Engineering and Computer Science, and Program in Survey Methodology.

MSSISS 2014 Keynote Address

Dr. Jeff Wu

Coca-Cola Chair in Engineering Statistics and Professor

Georgia Institute of Technology

From real world problems to esoteric research: some examples

Young (and some not-so-young) researchers often wonder how to extract good research ideas and develop useful methodologies from solving real world problems. The path is rarely straightforward and its success depends on the circumstances, tenacity and luck. I will use three examples to illustrate how I trod the path. The first involved an attempt to find optimal growth conditions for nano structures (i.e., wires, belts, saws). It led to the development of a new method "sequential minimum energy design (smed)", which exploits an analogy to potential energy of charged particles. After a few years of frustrated efforts and relentless pursuit, we realized that smed is more suitable for generating samples adaptively to mimic an arbitrary distribution rather than for optimization. The main objective of the second example was to build an efficient statistical emulator based on finite element simulation results with two mesh densities in cast foundry operations. It eventually led to the development of a class of nonstationary Gaussian process models that can be used to connect simulation data of different precisions and speeds. The third example hails from cell biology. In a T cell adhesion experiment at Georgia Tech, the biologist was not satisfied with the use of graphical method to understand the serial dependency of cell adhesion over repeated trials. It led to the development of hidden Markov models with new features that reflect the nature of the experiment. In each example, the developed methodology has broader applications beyond the original problem. I will explain the thought process in each example but cannot promise any general observation.

MSSISS 2014 Student Committee
Douglas Lehmann, Department of Biostatistics
Naveen Narisetty, Department of Statistics
Brandon Oselio, Department of Electrical Engineering and Computer Science
Brian Wells, Program in Survey Methodology

Faculty Advisory Committee
Timothy Johnson, Department of Biostatistics
Susan Murphy, Department of Statistics
Clayton Scott, Department of Electrical Engineering and Computer Science
Brady West, Program in Survey Methodology

For additional information, visit: http://sitemaker.umich.edu/mssiss/home

Student Presentation Table of Contents

Oral Presentation Session I 9:00 AM – 10:15 AM

A statistical decision procedure for personalizing treatment

Huitian Lei, Statistics, ehlei@umich.edu

In personalized treatment the recommended treatment is based on patient characteristics. Given pre-specified subgroups, we define the subgroup indicator as useful in personalized decision making if for particular subgroups there is sufficient evidence to recommend one treatment, while for other subgroups, either there is sufficient evidence to recommend a different treatment, or there is insufficient evidence to recommend a particular treatment. We propose a two-stage decision procedure to evaluate if a subgroup indicator is useful in personalized decision making. In the first stage of the procedure, we utilize the test statistic for testing treatment-subgroup interaction. If the first stage test statistic exceeds the critical value, we proceed to the second stage of the procedure and utilize test statistics for testing subgroup treatment effects. We choose the critical values to control the probability of making severe errors. We illustrate the proposed procedure using data from Child/Adolescent Anxiety Multimodal Study (CAMS).

Semiparametric methods to contrast restricted mean gap times

Xu Shu & Douglas E. Schaubel, Biostatistics, shuxu@umich.edu

Times between successive events (i.e., gap times) are of great importance in survival analysis. Very few existing methods allow for comparisons between gap times. Motivated by the comparison of primary and repeat transplantation, our interest is specifically in contrasting the gap time survival functions. Two major challenges in gap time analysis are non-identifiability of the marginal distributions and the existence of dependent censoring (for all but the first gap time). We use Cox regression to estimate the (conditional) survival distributions of each gap time (given the previous gap times). Combining fitted survival functions based on those models, along with multiple imputation applied to censored gap times, we then contrast the first and second gap times with respect to survival and restricted mean lifetime. Large-sample properties are derived, with simulation studies carried out to evaluate finite-sample properties. We apply the proposed methods to kidney transplant data obtained from a national registry.

Estimating a change-point in high-dimensional Ising model

Sandipan Roy, Statistics, sandipan@umich.edu

We present here a change-point model in the context of a high dimensional dynamic network. We discuss the estimation of the change-point when a high dimensional sparse network changes its structure at a particular point in time and subsequently estimate the network structures before and after the change-point. We propose a two stage estimation procedure involving l_1 penalized pseudo-likelihood method by Hoeffling and Tibshirani (2009) coupled with a nonparametric kernel smoothing procedure to estimate the change-point. We then solve two separate convex optimization problems to estimate the network structures before and after the estimated change-point. We show that the proposed method leads to consistent change-point estimation under high dimensional asymptotics. The proposed method is used to explore voting patterns in the US Congress and performance of the methodology for simulated data set is also presented.

Oral Presentation Session II 11:15 AM – 12:30 PM

Anatomy of network-level malicious activities: Connectedness and inter-dependence

Yang Liu, Electric Engineering and Computer Science, youngliu@umich.edu

We investigate the relationship between dynamic malicious activities of different network entities by using a set of reputation black lists (RBLs) collected over the past year. These RBLs capture the IP-level malicious activities in three broad categories: spam, phishing, and active scanning. We aggregate the RBL data at the prefix level and examine how the dynamic behavior of a prefix (measured by its presence on these RBLs) is similar to those of others. Furthermore, we explore to what extent such similarities are correlated with a number of spatial relationships between these networks, including their AS membership, AS type association, topological distances, as well as country affiliations. Our main findings are two-fold: (1) We show that a simple *similarity* measure that quantifies the relationship between two networks' dynamic behavior can capture hidden features previously unseen when only time-averaged quantities are used. (2) We then use statistical inference methods to evaluate the significance of the set of spatial features in explaining the observed similarity. Our results suggest that the topological distance between two networks is the most significant indicator of their similarity in maliciousness. We discuss various implications of these findings, and how they can be applied to better policy design.

Pre-surgical fMRI data analysis using a spatially adaptive conditionally autoregressive model

Zhuqing Liu, Biostatistics, zhuqingl@umich.edu

Spatial smoothing is an essential step in the analysis of functional magnetic resonance imaging (fMRI). One standard smoothing method is to convolve the image data with an isotropic Gaussian kernel, which applies a fixed amount of smoothing to the entire image. In pre-surgical brain image analysis where spatial accuracy is paramount, this method, however, is not reasonable as it may cause some regions to be undersmoothed while others oversmoothed, smearing out the boundaries of activation and deactivation regions of the brain. Moreover, in a standard fMRI analysis, strict false positive control is desired; for pre-surgical planning, false negatives are of greater concern. To this end, we propose a novel spatially adaptive conditionally autoregressive model with smoothing variances proportional to error variances, allowing the degree of smoothing to vary across the brain and present a new loss function, allowing for the asymmetric treatment of false positives and false negatives. We compare our proposed model with two existing spatially adaptive models and with a Bayesian non-parametric Potts model. Simulations studies show our model outperforms these other models; as a real model application, we apply the proposed model to a single subject's pre-surgical fMRI data to assess peri-tumoral brain activation.

Inference for monotone trend under dependence

Pramita Bagchi, Statistics, pramita@umich.edu

We consider the problem of estimating a monotone trend in a non-parametric regression setup with stationary dependent noise. We have studied this problem under both short-range and long-range dependence regimes for the noise sequence. In each case, we have proposed new efficient methods for constructing a point-wise confidence interval for the monotone trend function. Our methods are based on the method of inversion of test statistics that avoid the estimation of the derivative of the trend function, which is a common shortcoming of the existing Wald-type confidence intervals. We studied the asymptotic behavior of two types of statistics, L_n and Ψ_n. The first is an analog of a log-likelihood ratio, successfully used to tackle the same problem in the iid context, while the second is a new type of non-linear transformation of L_n and another discrepancy statistic. It is surprising that the asymptotic distribution of the statistic Ψ_n is essentially free of nuisance parameters in both the short- and long-range dependence regimes. In the latter, it involves only the Hurst long-range dependence parameter. Extensive simulations show that, in practice, our methods are robust to function shapes and various short-range dependence structures in the errors while providing intervals with accurate coverage.

Oral Presentation Session III 1:45 PM – 3:00 PM

Network reconstruction using nonparametric additive ODE models

James Henderson & George Michailidis, Statistics, jbhender@umich.edu

Network representations of biological systems are widespread and reconstructing unknown networks from data is a focal problem for computational biologists. For example, the series of biochemical reactions in a metabolic pathway can be represented as a network, with nodes corresponding to metabolites and edges linking reactants to products. In a different context, regulatory relationships among genes are commonly represented as directed networks with edges pointing from influential genes to their targets. Reconstructing such networks from data is a challenging problem receiving much attention in the literature. There is a particular need for approaches tailored to time-series data and not reliant on direct intervention experiments, as the former are often more readily available. In this paper, we introduce an approach to reconstructing directed networks based on dynamic systems models. Our approach generalizes commonly used ODE models based on linear or nonlinear dynamics by extending the functional class for the functions involved from parametric to nonparametric models. Concomitantly we limit the complexity by imposing an additive structure on the estimated slope functions. Thus the submodel associated with each node is a sum of univariate functions. These univariate component functions form the basis for a novel coupling metric that we define in order to quantify the strength of proposed relationships and hence rank potential edges. We show the utility of the method by reconstructing networks using simulated data from computational models for the glycolytic pathway of *Lactocaccus Lactis* and a gene network regulating the pluripotency of mouse embryonic stem cells. For purposes of comparison, we also assess reconstruction performance using gene networks from the DREAM challenges. We compare our method to those that similarly rely on dynamic systems models and use the results to attempt to disentangle the distinct roles of linearity, sparsity, and derivative estimation.

Item nonresponse in auxiliary variables used in calibration for survey sample data

Raphael Nishimura, Survey Methodology, raphaeln@umich.edu

Under the condition of absence of nonsampling errors, the most prominent statistical role of calibration methods in surveys is to reduce the sampling variability of the estimates. This is achieved when the survey variables are associated with the auxiliary variables used in the calibration procedure. Furthermore, in the presence of undercoverage and nonresponse, these methods can potentially reduce bias caused by these sources of error. Calibration, how-ever, assumes that the auxiliary variables are completely observed for all the responding units. When there is a high missing rate in one or more auxiliary variables, in practice, these variables are usually not used in calibration, even if they are important to explain the survey variables. Very few papers have analyzed this problem and none of them have considered imputation as a possible solution. Hence, it is important to evaluate the performance of such calibration estimators when imputation is used to fill in these missing data on the auxiliary variables used in the weighting adjustments. Moreover, it is important to examine how well sampling variance is estimated with multiple imputation in this case. In this paper it is shown the results of a simulation study conducted to compare the Horvitz-Thompson estimator with a special case of calibration, the Generalized Regression (GREG) estimator, in the presence of missing data in one of two variables used in the calibration adjustment and multiple imputation is used to fill them in.

Adaptive phase I clinical trial design using Markov models for conditional probability of toxicity in oncology

Laura Fernandes, Jeremy M.G. Taylor, & Susan Murray, Biostatistics, flaura@umich.edu

Many cancer therapies in phase I trials in oncology involve multiple dose administrations to the same patient over multiple cycles, with a typical cycle lasting three weeks and having about six cycles per patient. The goal of these studies includes finding the maximum tolerated dose (MTD) and studying the dose toxicity relationship. The dose a patient receives is usually not changed from one cycle to the next. Most studies currently reduce the data to a binary end point, the occurrence of a toxicity and analyze the data either by considering the toxicity from the first dose or from any cycle on the study when the toxicity occurred. This paper provides an alternative approach that would allow the dose for a patient to change from one cycle to the next and to gather the toxicity data from each cycle. This extra information from each cycle provides more precise estimation of the dose-toxicity relationship and enables a better selection of the dose for the next patient and at the conclusion of the trial. A Markov model is formulated for the conditional probability of toxicity on any cycle given no toxicity in previous cycles and given the current and previous doses. Simulation results are presented demonstrating the efficiency gains in using the new Markov model compared to analyses of a single binary outcome. The method is extended to running an adaptive trial with sequential dose assignment to patients completing a cycle without a dose limiting toxicity incorporating all the information until that point.

Poster Session I 10:15 AM – 11:15 AM

Incorporating Expected Utility and Contingency Plans in Kidney Paired-Donation Optimization Schemes

Mathieu Bray, Biostatistics, braymath@umich.edu

In a KPD pool, pairs consisting of a transplant candidate and their recruited incompatible donor are matched with other complementary pairs and altruistic donors (AD) in an attempt to find combinations such that all recipients involved can obtain a transplant. This simulation study considers optimization schemes for determining the ideal transplant arrangement based on maximum utility and expected utility, as well as sets and components of pairs producing the maximum total expected utility amongst all embedded sub-cycles and sub-chains (contingency plans). Data consists of pairs and ADs from the Alliance for Paired Donation as well as the Michigan KPD program. 200 Monte Carlo simulations were performed for each optimization scheme in a dynamic KPD pool over 8 months, with 30 pairs and 1 AD added to the pool per month. A virtual crossmatch was performed for every possible transplant between donor and recipient, and utilities and probabilities for each transplant were generated, representing the relative worth and the probability of failure respectively. At each match run, the optimal set of cycles and chains was determined based on the criteria of the scheme in question. Failures (on both transplant and pairs) were simulated and the realized utilities of these transplants were reported. We find that on average, up to 20% more transplants are realized by both taking into account failure probabilities and allowing contingency plans, as opposed to simply maximizing utility. This suggests there is merit to determining the optimal arrangement based on expected utility amongst all possible contingency plans, as opposed to simply maximizing the possible utility, either outright or post hoc.

Modeling Student Course Grades with Random Forest Regression

Omar Chavez, Statistics, odchavez@umich.edu

A predictive probability model of student academic achievement, measured by final course grade was developed to give students and educators real time feedback of student progress for University of Michigan's STATS 250 course during Winter 2014. We used a Random Forest regression algorithm where the response variable was Final Curse Grade on a scale from 0 to 100% and predictor variables that included GPA at the start of term, any homework and exam grades available at the time the model was displayed, as well as several behavior variables including average amount of sleep per week, what academic resources were utilized by students (office hours, tutors, textbook, class notes, etc) and finally, the student's score from Felder's Index of Learning Styles Survey. The training data consisted of approximately 1700 observations from Fall 2013 and was collected from course grades and weekly surveys administered to students via homework. The motivation for sharing the grade model with students is to help students set realistic achievement goals that might otherwise seem out of reach as well as to offer suggestions that can help students achieve their individual goal grades if their current expected grade is inconsistent with their desired grade. We also use it as a tool to inform educators of at-risk students that might need extra attention as well as to identify students that are performing much better than would otherwise be expected based on what we know about the student. Our motivation for identifying such students is to mine the population of students for effective study habits that are most effective in STATS 250 when one controls for available information listed above.

Enhanced Peters-Belson: Bias in treatment-prognostic score interaction

Josh Errickson, Statistics, jerrick@umich.edu

Recent papers in empirical economics (Giné et al. (2012), Dynarski et al. (2011), and others) have been interested in examining if a treatment has more effect among those who are predicted to have the worst response in the absence of any treatment. The statistical validity of this method has not been verified. There has been some work showing simulationally that this method introduces bias into the estimate of the interaction between predicted response and treatment effect. We examine this method more rigorously, and show a corrected standard error estimate, as an attempt to analyze the effect of the bias in large sample problems.

A Study of Bootstrap inference on LASSO-type estimators in a moving-parameter perspective

Jun Guo & Stephen M.S. Lee, Statistics, guojun@umich.edu

Growing interest in LASSO-type estimators has brought up an important question of how their distributions, often complex and nontrivial, can be consistently estimated. Success in doing so will clear the way for us to make valid statistical inference based on LASSO-type estimators. Working within the traditional, fixed-parameter, asymptotic framework, theoretical studies have been done on various bootstrap methods to identify successful solutions to the problem. However, recent empirical studies have shown that the “successful” methods, when put into practice, do not necessarily outperform, and are sometimes even inferior to, the “unsuccessful” ones. The apparent incompatibility between asymptotic and empirical findings casts doubt upon the very usefulness of the traditional fixed-parameter framework in studies of this kind. A new, moving-parameter, asymptotic framework has very recently been advocated to provide a broader scope for investigating asymptotic properties of statistical quantities and hence a better explanation of finite-sample results. This research aims to provide a thorough re-assessment of bootstrap strategies in this new moving-parameter framework for making LASSO-based inference. We hope that the new findings can be utilized to make general, practically more relevant recommendations on the bootstrap method best suited to estimate the distribution of the particular LASSO-type estimator in question.

Adaptive mobile phone interventions for the treatment of substance abuse: A study of feasibility, the current state of the art, and possibilities for more effective, data-driven tailoring

Kelly Hall, Zach Murray, Abigail Sagher, Shawna Smith, & Susan Murphy, Statistics, kellyhal@umich.edu

Our lab studies and develops mobile phone applications that use data and algorithms to provide tailored support to at-risk individuals with health problems. For certain hard-to-reach populations (i.e. the homeless or those of low SES), traditional treatment for substance use disorders is not often easily available. Further, treatment in a clinical setting cannot always prepare and protect individuals for challenges their daily lives. The moments when they are most vulnerable – right after work, walking past their favorite liquor stores – are the moments when they truly need support. The young, rapidly-evolving field of adaptive mobile phone interventions might provide the solution.

Mobile phone applications are being developed to provide support at the moments when relapse is likely – the moments when users need help most. Several studies have shown the feasibility of administering health support to this vulnerable population via mobile phones, and applications have already been created using a variety of successful treatment components, from cognitive reframing strategies to social support. As technology has advanced, the state of the art has evolved from basic, non-personalized treatment (i.e. sending encouraging text messages three times per day) to baseline, theory-driven tailoring (i.e. sending warnings at the times of day users have indicated to be the most challenging).

The future of the field looks bright, and statisticians can play a critical role. As tailoring becomes increasingly advanced and data-driven, algorithms can be developed to determine, for example, what situations or self-reported emotions indicate eminent cravings, how many times per day users prefer to be prompted, and which intervention strategies are most effective. Algorithms could allow the application to adapt to the user, learning his or her patterns and triggers and adjusting for maximum effectiveness. For vulnerable populations struggling with substance use disorders, these adaptive interventions could be economical, widely adopted, individualized, and even life-changing.

Strengthening an instrumental variable when matching

Doug Lehmann, Yun Li, & Yi Li, Biostatistics, lehmannd@umich.edu

Instrumental variable (IV) methods are commonly used when there is a concern that unobserved confounders exist between the treatment and outcome. This is almost always a concern in observational studies where the researcher does not have control over treatment assignment. Near/far matching has been proposed as an IV methodology for use with binary outcomes that aims to pair subjects that appear similar (near) on relevant covariates, but different (far) in the levels of encouragement toward treatment each received. Increasing the average difference of encouragement within pairs creates a stronger instrument, which in turn leads to results that are less sensitive to violations of key IV assumptions. We compare two approaches to strengthening the instrument in the near/far matching framework. The first weights pairs based on the within-pair instrument strength, while the second optimally removes a portion of subjects that received only moderate encouragement, as opposed to strong encouraged toward treatment or control, from the analysis. While both successfully create a stronger instrument, they differ in their effect on the resulting match quality. The strengths and weaknesses of each method are investigated through simulation studies, and practical guidance for selecting which is appropriate in various scenarios is discussed. Both methods are illustrated using Medicare data to examine the relationship between treatment modality and outcomes for patients receiving dialysis as treatment for end stage renal disease.

Group Comparison of Pulsatile Hormone Times Series

TingTing Lu & Timothy Johnson, Biostatistics, ttlu@umich.edu

Due to its oscillatory and pulsatile nature, analyzing hormone time series data is challenging and many model-based methods have been proposed over the years. Typically, analyses are performed in two stages. First, the number and locations of the episodic events are determined. Second, a model is fit to the data conditional on the number of pulses. However, errors occurring in the first step are carried over to second. In, 2007, Johnson proposed the first fully Bayesian deconvolution model that jointly estimates both the number and locations of secretion events and admits a non-constant basal concentration. Thus both pulsatile and oscillatory components of hormone secretion are simultaneously modeled. Furthermore, the model allows for variation in pulse shape and size. However, the model cannot handle groups of subject and cannot compare secretion patterns between groups of subjects. In this paper we extend Johnson's model in two ways. First, we admit group comparisons of the underlying pulse driving mechanism. Second, we model the pulse driving mechanism via a Cox process where the intensity function is not assumed constant as is assumed in Johnson (2007). We take a fully Bayesian hierarchical approach to estimate model parameters. We apply our new modeling approach to a hormone study of depressed women with age matched healthy controls and compare results with a smoothing spline functional data analysis approach.

A hybrid second order iterated smoother

Dao Nguyen, Statistics, nguyenxd@umich.edu

Plug and play inferences, also known as derivative free or likelihood free inferences are receiving great attention recently due to the fact that in many practical problems, the likelihood is intractable to compute directly. The attractive properties of these methods are that unobserved process enters the algorithm only through the requirement that realizations can be generated at arbitrary parameter values. In the same line with plug and play approach, this paper introduces a hybrid second-order iterated smoother. To reduce the high variance of these estimators, we use fixed lag smoother. To improve the speed of convergence, an approximation of the observed information matrix is also proposed. While enjoying greater convergence rate, most observed information matrix approximation are computational expensive, especially in plug and play approaches. Therefore, to relax the intensive computation, we use Newey-West covariance estimator for a few initial iterations before adapting sequential Monte Carlo approximations of the variance. Due to the special structure of iterated smoothing, we also bypass the sequential Monte Carlo approximations of the variance by using the last estimated smoother value at the beginning of the next filter iteration. In a toy example and in a challenging inference problem of fitting a malaria transmission model to time series data, we find substantial gains for our methods over current alternatives.

Bivariate Interaction Models

Karen Nielsen, Statistics, karenen@umich.edu

When exploring data with possible interactions between predictor variables, researchers often artificially dichotomize or split one continuous variable. This allows for simple two-dimensional plots and tests of slope for specified levels of the predictors, but it oversimplifies the situation. Some work has already been done on probing the simple interaction between two continuous variables (e.g., Bauer and Curran, 2005), but even that may not be enough to capture the complex relationship that might exist between predictors. We propose a bivariate 2nd order interaction model in the context of the generalized linear mixed model for exploring the interaction between variables and the resulting effects on the response. This approach also allows a novel testing method by focusing on an underlying basis set for each variable polynomial or otherwise. Potential interpretation advantages this may yield and the importance of using informative visualizations to guide inference will be covered.

Inference for disease dynamics in multiple cities using Sequential Monte Carlo: A Case Study in Measles

Joonha Park, Statistics, joonhap@umich.edu

Simulation of nonlinear dynamic systems of high dimension has been regarded as computationally heavy problems. As a result, little efforts have been made to jointly estimate disease epidemics in multiple cities. In this report, we propose a variation of Sequential Monte Carlo (SMC) method for estimating latent states which can provide a computationally feasible solution to the joint estimation of dynamic trajectories of interacting systems. This method, which we name Factorial Particle Filter (FPF), is expected to reduce the computational cost by a huge amount when the sub-systems are weakly interacting, while achieving the desired property that the sampled latent states form a proper sample from its true distribution according to the underlying model. We apply this method to the measles epidemic in the UK from 1950 to 1953. The analysis performed on the five largest cities by population shows that the proposed Factorial Particle Filter method yields a reasonable estimate of epidemic history with relatively low computational cost. The conventional SMC method applied on the same set of data could not generate any result. We also estimate key epidemic parameters using the Iterated Filtering method (Ionides et. al, 2011), and show that the transmission rate is substantially different between during school term and during school holidays.

On a fused lasso approach to combining parameter estimates with multiple imputation data

Catherine Robertson, Lu Tang, Wen Wang, Lu Xia, & Peter X.K. Song, Biostatistics, ccrober@umich.edu

In statistical analysis with missing values, traditionally multiple imputation methods simply take the average of parameter estimates generated with each imputed set. In this paper we propose a new method of combining parameter estimates via the fused lasso approach. It turns out that the conventional average-based method is a special case of our new method when the tuning parameter of the fused lasso penalty is set as zero. We establish the procedure of estimate combination within the framework of generalized lasso, for which the statistical software is already ready to be applied. Using extensive simulation studies, we demonstrate the performance of our new method, including the comparison to the conventional method in terms of estimation bias and prediction power, where both parametric multiple imputation and nonparametric imputation are considered. A real biomedical data analysis example is also provided for illustration.

Sequential Stratification for Recurrent Event Outcomes

Abigail Smith & Douglas Schaubel, Biostatistics, abbysmit@umich.edu

Recurrent events are of increasing interest in observational studies, and in some of these studies the goal is to estimate the effect of a certain treatment on the recurrent event rate. If two or more treatments could potentially occur, but only the first is observed, the ideal comparison is between the treatment of interest and any other potential treatment course. Sequential stratification is a method for estimating the effect of choosing one treatment course relative to waiting and potentially receiving another treatment course on terminal events such as death. Since the method has only been developed for terminal events such as death, we extend sequential stratification to the recurrent event setting. The objective of this analysis is to extend this method to the recurrent event setting. Asymptotic properties of the estimators and variances are explored. The performance of the method in moderate sized samples is assessed through simulation. Finally, the proposed method is applied in a clinical dataset to evaluate the effect of living donor liver transplantation on hospitalization rates.

Inference on Infectious Disease Dynamics from Viral Genetic Sequences

Alex Smith, Bioinformatics, alxsmth@umich.edu

The rapid evolution of RNA viruses may provide a means of understanding fundamental mechanisms—and potential points of control—that shape the course of the epidemics they cause. In particular, because viral evolution and disease dynamics occur on similar timescales, when an epidemic moves through a host population it leaves a signature in the genetics of the pathogen population. Genetic sequences of pathogens sampled through time may therefore contain information on fundamental mechanisms that drive disease dynamics. Recent advances in sequencing have increased the quality and availability of viral genetic sequences. However, current methods for inference on dynamics from sequence data are limited both in the classes of models they can consider and in the size and nature of datasets they can analyze. We have developed a framework to fit dynamic disease models to both longitudinal genetic data on pathogens and epidemiological data on hosts. The key novelty of our methodology is coupling phylogenetic techniques with partially observed Markov process models of disease dynamics. In doing so, we have created a set of algorithms with the flexibility to assess a diverse set of models that correspond to specific mechanistic hypotheses and thus opened a rich and growing data type to new lines of scientific inquiry. Here, I outline the scope of the inference problem, describe the essential features of our approach and present preliminary results from an analysis of simulated data.

Improving Trauma Triage Models for Motor Vehicle Crashes

Yaoyuan Vincent Tan, Biostatistics, vincetan@umich.edu

Severe injury prediction in motor vehicular crashes has long been an interest of trauma researchers. Many studies have shown that delta-v, a measure of the near-instantaneous change in velocity during impact of a crash, is a strong predictor of severe injuries. Delta-v is often estimated by re-constructing the accident after crash investigations. However, this process has been shown to underestimate delta-v. Recent advances in technology have enabled a device, the Event Data Recorders (EDR), to capture the full deceleration trajectory during crashes when air-bags are deployed. We propose using information from these deceleration trajectories to estimate risk of injury. In particular, we propose to use functional data analysis (FDA) to estimate the mean trend in a deceleration profile. We then obtain integrals of these mean trends as an estimation of delta-v, along with integrals of the absolute value of the derivate of delta-v, as well as the residuals variance. We then use these three elements as key summary measures of deceleration to predict injury. We apply our method to 2005-2011 EDR data sets available on the National Highway and Transportation Safety Administration (NHTSA) website. We enhance our prediction model with baseline covariates found in National Automotive Sampling System (NASS) Crashworthiness Data System (CDS). We show in our results that the key information we summarize from the longitudinal deceleration trajectories enhances prediction relative to that obtained using baseline characteristics only. Because information from the EDRs can be transmitted to Emergency Medical Services (EMS) immediately after a crash, our results will be useful to predict severe injuries more accurately, allocate resources more efficiently, and ultimately reduce morbidity and mortality in passenger vehicle crashes.

A comparison of MCMC and Variational Bayes Algorithms for Log-Gaussian Cox Processes

Ming Teng, Biostatistics, fcxtm355@gmail.com

Log-Gaussian Cox Processes (LGCP) are flexible models for fitting spatial point pattern data. In order to estimate the intensity function, a Bayesian model with implementation based on Markov chain Monte Carlo (MCMC) simulation from the posterior, proposed by Møller et al. (1998), is commonly used. However, for LGCPs, MCMC is slow to converge to the posterior distribution and mixing is slow thereafter. Møller et al. proposed the use of the Metropolis adjusted Langevin algorithm (MALA), which helps considerably with mixing. More recently, Coeurjolly and Møller (2013) proposed a Variational estimator for LGCPs. We consider both MALA and variational Bayes methods based on mean field approximations for fitting LGCP models to 3D point pattern data with subject specific covariates and spatially varying coefficients. The application of VB to LGCP models is made challenging by the non---conjugate structure of the model. To develop tractable solutions, we incorporate Laplace approximations within the VB framework (Wang and Blei, 2013) which leads to Gaussian variational approximations. We make comparison between MALA and VB in terms of statistical and computation efficiency. Simulation studies are used to evaluate efficiency and we apply the algorithms to an imaging study of Multiple Sclerosis lesion locations with subject specific covariates.

Community Detection in Networks with Node Features

Yuan Zhang, Elizaveta Levina, & Ji Zhu, Statistics, yzhanghf@umich.edu

Many methods have been proposed for community detection in networks, but most of them do not take into account additional information on the nodes that is often available in practice. In this paper, we propose a new joint community detection criterion that uses both the network and the features to detect community structure. One advantage our method has over existing joint detection approaches is the flexibility of learning the impact of different features which may differ across communities. Another advantage is the flexibility of choosing the amount of influence the feature information has on communities. The method is asymptotically consistent under the block model with additional assumptions on the feature distributions, and performs well on simulated and real networks.

Poster Session II 12:30 PM – 1:45 PM

Predicting Participation in a Mobile-Web Survey

Christopher Antoun, Survey Methodology, antoun@umich.edu

A growing number of people are responding to Web surveys on their phones. This emerging method of survey data collection, which is commonly referred to as "mobile Web," provides new opportunities for survey researchers. For example, by enabling respondents to complete surveys where it is convenient, mobile-Web surveys could have higher participation than other modes. Yet, early research suggests that people are less willing to participate in mobile-Web surveys than traditional Web surveys. Why is this? The current study explores the mechanisms, or causes, of nonresponse in a mobile-Web survey. 1263 members of the Longitudinal Internet Studies for the Social Sciences (LISS) panel completed a traditional Web survey about attitudes toward and use of technology. Next, these same panel members were invited to participate in a mobile-Web survey, and we observed who responded and who did not. Nonresponse propensity models were developed using both demographic characteristics from the LISS panel frame and technology measures from our baseline survey. We explore the correlates of nonresponse. Even after controlling for demographic differences in response rates, we expect to find that technology measures, such as type of phone and frequency of mobile-Internet usage, will predict participation. These findings will help add to an emerging picture of why people do or do not participate in Web surveys using their phones.

Regularized Estimation in Sparse High-dimensional Time Series Models

Sumanta Basu, Statistics, sumbose@umich.edu

Many scientific and economic problems require the analysis of high-dimensional time series datasets. However, theoretical studies in high-dimensional statistics to date rely primarily on the assumption of independent and identically distributed (i.i.d.) samples. In this work, we investigate the theoretical properties of ℓ_1-regularized estimates in three important statistical problems in the context of high-dimensional time series - (a) stochastic regression with serially correlated errors, (b) transition matrix estimation in vector autoregressive (VAR) models, and (c) covariance matrix estimation from temporal data. For all three problems, we derive non-asymptotic upper bounds on the estimation errors, thus establishing that consistent estimation is possible via ℓ_1-regularization for a large class of stationary time series under sparsity constraints. A key technical contribution of the work is to introduce a measure of stability for stationary processes, that provides insight into the effect of dependence on the accuracy of the regularized estimates. Further, we establish some useful deviation bounds for statistics generated from dependent data, which are of independent interest.

Perceptron-like Algorithms and Generalization Bounds for Learning to Rank

Sougata Chaudhuri, Statistics, sougata@umich.edu

Learning to rank is a supervised learning problem where the output space is the space of rankings but the supervision space is the space of relevance scores. We make theoretical contributions to the learning to rank problem both in the online and batch settings. First, we propose a *perceptron*-like algorithm for learning a ranking function in an online setting. Our algorithm is an extension of the classic perceptron algorithm for the classification problem. Second, in the setting of batch learning, we introduce a *sufficient condition* for convex ranking surrogates to ensure a generalization bound that is independent of number of objects per query. Our bound holds when linear ranking functions are used: a common practice in many learning to rank algorithms. En route to developing the online algorithm and generalization bound, we propose a novel family of *listwise* large margin ranking surrogates. Our novel surrogate family is obtained by modifying a well-known *pairwise* large margin ranking surrogate and is distinct from the listwise large margin surrogates developed using the structured prediction framework. Using the proposed family, we provide a guaranteed upper bound on the cumulative NDCG (or MAP) induced loss under the perceptron-like algorithm. We also show that the novel surrogates satisfy the generalization bound condition.

Sparse Approximation of Kernel Means

Efren Cruz-Cortes, Electrical Engineering and Computer Science, encc@umich.edu

We examine the problem of approximating the mean of a set of vectors by a sparse linear combination of these vectors. Our motivation springs from common machine learning problems, where a probability distribution can be estimated by the sample mean of kernel functions. Scalability is essential for cases with large sample sizes or when the kernel function mean has to be evaluated repeatedly, therefore existing algorithms for sparse approximation, such as matching and basis pursuit, are not good candidates for these problems. We introduce an approximation bound involving a novel incoherence measure, and propose bound minimization as a sparse approximation strategy. In the context of sparsely approximating a kernel mean function, the bound is efficiently minimized by solving an appropriate instance of the k-center problem, and the resulting algorithm has linear complexity in the sample size.

Off-line Off-policy reinforcement learning

Yanzhen Deng, Statistics, dengyz@umich.edu

In this work we develop a data analysis method for constructing a treatment policy. This treatment policy would be used by a smartphone to deliver behavioral interventions. The data analysis method is an off-line off-policy reinforcement learning method. We want to learn a treatment policy that will maximize the average of a longitudinal outcome (e.g. the "reward"); this average is called the average reward. Our method can be used with a data set in which the treatments are assigned by a known policy. This data set will be composed of n trajectories, each of the form $S_1, A_1, R_2, S_2, A_2, R_3, \ldots R_T, S_T$, where S is covariates of the subject, A is the treatment, and R is the reward. Our method solves estimating equations based on the Bellman equation. Although the average reward is a long-term goal, we don't require long trajectory for the subjects. Initial simulations show that the average reward is well estimated and that our method leads to an optimal treatment policy.

Efficient Estimation of Partial Rank-based Correlation with Missing Data

Wei Ding, Biostatistics, dingwei@umich.edu

Rank-based correlation is widely used in practice to measure dependence between variables when their marginal distributions are skewed. Estimation of such correlation is challenged by both the presence of missing data and the need of adjusting for confounders. In this paper, we develop a unified framework of Gaussian copula regression that enables us to estimate both partial Pearson correlation and partial rank-based correlation (e.g. partial Kendall's tau or partial Spearman's rho), depending on the assumed marginal distributions. To adjust for confounding factors, we utilize marginal regression models with location-scale distributions for error terms. We establish the EM algorithm in this semi-parametric modeling framework to handle the estimation with missing values. The semi-parametric efficiency of our estimation method is also discussed. We propose a peeling procedure to carry out iterations required in the EM algorithm. We compare the performance of our proposed method to the traditional multiple imputation approach through simulation studies. For structured correlation, such as an exchangeable or a first-order auto-regressive (AR-1) correlation, our method outperforms multiple imputation approach in terms of both bias and efficiency.

Controlling a Random Network with Linear Dynamics

Mohamad Kazem Shirani Faradonbeh, Statistics, shirany@umich.edu

We consider a random network whose nodes can be in different states taking values in a real-valued set. Further, the network evolves over time according to linear dynamics. The problem under study is how to steer the network to "desirable" states. This problem captures key features of applications in social network analysis, marketing science and engineering communication networks. To achieve the posited goal, a subset of nodes can act as controllers. We present algorithms on how to select a minimum number of controllers. Further, we establish results on trade-offs between selection efficiency and computational complexity of the algorithms and examine how these issues are affected by the structure of the underlying random network topology.

Multi-marker tests for joint association in longitudinal studies using the genetic random field model

Zihuai He, Biostatistics, zihuai@umich.edu

Longitudinal genetic studies of common and chronic diseases risk factors provide a valuable opportunity to explore how genetic variants affect traits over time by utilizing the full trajectory of longitudinal outcomes. Since disease risk factors and phenotypes are likely influenced by the joint effect of multiple variants in a gene, a joint analysis of these variants considering linkage disequilibrium (LD) and potential interactions among the variants may also help to explain additional heritability. In this article, we propose a longitudinal genetic random field model (LGRF), to test the joint association between a set of genetic variants and a phenotype measured repeatedly during the course of an observational study. The phenotypes of subjects are modeled as a random field on a space spanned by their sequenced genotypes and time. Generalized score tests are developed for testing the joint association and their asymptotic properties are derived. Several essential methodological improvements necessary for improving robustness to the misspecification of within-subject correlation structure and scalable implementation in large-scale GWAS are further proposed. The proposed methods are evaluated through extensive simulation studies and illustrated using data from the Multi-Ethnic Study of Atherosclerosis (MESA). Our simulation results indicate substantial gain in power using LGRF when compared to the two commonly used existing alternatives: (i) single marker tests using longitudinal outcome and (ii) existing multi-marker association tests such as the sequence kernel association tests (SKAT) using the average value of repeated measurements as the outcome.

Analyzing complex survey data covering campus sustainability issues

Qiaoxian Hu, Survey Methodology, qxhu@umich.edu

The Sustainability Cultural Indicators Program (SCIP) at the University of Michigan is an annual Web survey designed to measure and track the behavior, knowledge, and culture of students, faculty, and staff at Ann Arbor campus. The survey, based on samples of over 4000 students and 2000 faculty and staff was first conducted in the fall of 2012. SPSS Complex Sample Analysis Module was used in analyzing the data. Unlike other software packages for survey data analysis, a Complex Sample Analysis plan has, to be created in SPSS before conducting complex survey data analysis. Failure to build the Complex Sample Analysis plan will lead to less accurate variance estimations. The SCIP 2012 analysis incorporates sample weights, which adjust for gender, status (faculty, staff, student grade level), and whether in the Health System or not (for faculty and staff only) and strata. This study will introduce how SPSS Complex Sample Analysis Module works in the SCIP analysis and will compare the estimations under the Base SPSS and the SPSS Complex Sample Analysis.

Semiparametric Approach for Regression with Covariate Subject to Limit of Detection

Shengchun Kong, Biostatistics, kongsc@umich.edu

We consider generalized linear regression analysis with left-censored covariate due to the lower limit of detection. The complete case analysis by eliminating observations with values below limit of detection yields valid estimates for regression coefficients, but loses efficiency. Substitution methods are biased; maximum likelihood method relies on parametric models for the unobservable tail probability, thus may suffer from model misspecification. To obtain robust and more efficient results, we propose a semiparametric likelihood-based approach for the regression parameters using an accelerated failure time model for the covariate subject to limit of detection. A two-stage estimation procedure is considered, where the conditional distribution of the covariate with limit of detection given other variables is estimated prior to maximizing the likelihood function for the regression parameters. The proposed method outperforms the complete case analysis and the substitution methods as well in simulation studies. Technical conditions for desirable asymptotic properties are provided.

Learning with Perturbations via Gaussian Smoothing

Chansoo Lee, Electrical Engineering and Computer Science, chansool@umich.edu

We study the regularization by stochastic perturbation in online linear optimization settings. Our novel analysis technique leverages a useful observation that perturbing data is equivalent to adding a regularization penalty on the decision space, allowing us to express regret in terms of the 2nd order behavior of the smoothed dual function. We extend the previous results on Gaussian smoothing and prove much tighter bounds for commonly used objective functions. We show that our perturbation via Gaussian smoothing framework produces low-regret algorithms for both the L1/L∞ setting (the "experts setting") and the L2=L2 setting. In each case, the worst-case regret is optimal to within small multiplicative constants, which is a significant improvement over previous results using perturbation techniques.

Classification of gene-gene associations using publicly available expression datasets

Qixing Liang, Biostatistics, liangqx@umich.edu

The Gene Expression Omnibus (GEO) is an online repository containing expression measurements of tens of thousands of genes in tens of thousands of samples. We hypothesize that co-expressed genes are likely to have functional relationships, and characterizing similarities across gene expression patterns may help predict gene-gene associations. We apply classifiers based on support vector machines and L1-penalized logistic regression models to predict gene-gene associations using the GEO dataset and also use predicted gene regulatory relationships from the hmChIP dataset to supervise our classifiers. These classifiers show improved performance compared to the naïve method of using correlation coefficients. The predicted gene-gene associations may be useful for better guiding downstream analyses, such as gene clustering, pathway or functional analyses.

Improved Robust PCA Using Low-rank Denoising with Optimal Singular Value Shrinkage

Brian Moore, Electrical Engineering and Computer Science, brimoor@umich.edu

We study the robust principal component analysis (PCA) problem of reliably recovering a low-rank signal matrix from a signal-plus-noise-plus-outliers matrix. We analytically characterize the extent to which the singular vectors of the signal-plus-noise-plus-outliers matrix can be degraded by outliers and discuss why a recently proposed method for robust PCA that exploits outlier sparsity to improve low-rank estimation will produce suboptimal low-rank matrix estimates in the presence of noise. Next, we propose a new iterative algorithm for robust PCA that utilizes an optimal, data-driven low-rank matrix estimator (OptShrink) derived in the context of random matrix theory. Finally, we show that the proposed approach yields superior background subtraction on a computer vision dataset.

A New Gibbs Sampler for Consistent Model Selection in High Dimensional Sparse Logistic Regression

Juan Shen, Statistics, shenjuan@umich.edu

We propose a Bayesian variable selection method for logistic regression that adapts to both the sample size n and the number of potential covariates p with two important features. First, we use spike and slab priors on the regression coefficients that shrink and diffuse, respectively, as the sample size increases. The shrinking and diffusing priors allow us to establish strong selection consistency even when $p > n$. Second, we propose a new Gibbs sampler that does not require p^2 operations in each of its iterations, but retains the property of strong posterior consistency. In contrast with the standard Gibbs sampler which requires sampling from a p dimensional multivariate normal distribution with a non-sparse covariance matrix, our new algorithm is much more scalable to high dimensional problems, both in memory and in computational efficiency. We compare our proposed method with several leading variable selection methods through a simulation study to show that our proposed approach selects the correct model with higher probabilities than most other methods.

A non-parametric approach for detecting differential alternative splicing in RNA-seq data

Yang Shi, Biostatistics, shyboy@umich.edu

High-throughput sequencing of transcriptomes (RNA-Seq) has rapidly evolved as a powerful tool for the study of gene expression and alternative splicing in humans and model organisms. With the reduction of the cost of sequencing, researchers are able to design complicated RNA-Seq experiments and generate large-scale RNA-seq data with hundreds or thousands of samples. We present a non-parametric approach to detect differential splicing of alternative isoforms for large sample RNA-Seq data. Our approach are more robust to outliers that often exist in large sample size RNA-seq experiments. Simulations studies show our approach has well-controlled type-I error rate and good power in detecting differential splicing events. We also compare our approach with other parametric methods in simulations and a real RNA-seq dataset from prostate cancer patients.

Multiple Decisions for Altruistic Donors in Kidney Paired Donation Program

Wen Wang, Biostatistics, wangwen@umich.edu

In contrast to growing need for kidney transplants to treat end-stage renal disease, supply of transplantable kidneys is in serious shortage. Kidney paired donation (KPD) program serving as a platform for candidates with willing but incompatible donors to exchange donors creatively increases number of available kidneys. Recently, altruistic donors (ADs) have been introduced into KPD pools by allocating an AD to an incompatible candidate-donor pair, whose candidate is compatible with the AD and whose donor agrees to donate to another pair in the pool, and so on. The sequence of transplants extended from the AD was arranged one transplant at a time in practice. This study addressed whether arranging multiple transplants for an AD was beneficial. Simulations based on medical records were preformed to investigate the impact of arranging multiple transplants for each AD on number of transplants, algorithm complexity and cost. Results suggested arranging multiple transplants for an AD significantly increased number of transplants given limited computational ability and cost.

Disease Prediction based on Functional Connectomes using a Scalable and Spatially-Informed Support Vector Machine

Takanori Watanabe, Electrical Engineering and Computer Science, takanori@umich.edu

Substantial evidence indicates that major psychiatric disorders are associated with distributed neural dysconnectivity, leading to strong interest in using neuroimaging methods to accurately predict disorder status. In this work, we are specifically interested in a multivariate approach that uses features derived from whole-brain resting state functional connectomes. However, functional connectomes reside in a high dimensional space, which complicates model interpretation and introduces numerous statistical and computational challenges. Traditional feature selection techniques are used to reduce data dimensionality, but are blind to the spatial structure of the connectomes. We propose a regularization framework where the 6-D structure of the functional connectome (defined by pairs of points in 3-D space) is explicitly taken into account via the fused Lasso or the GraphNet regularizer. Our method only restricts the loss function to be convex and margin-based, allowing non-differentiable loss functions such as the hinge-loss to be used. Using the fused Lasso or GraphNet regularizer with the hinge-loss leads to a structured sparse support vector machine (SVM) with embedded feature selection. We introduce a novel efficient optimization algorithm based on augmented Lagrangian and the classical alternating direction method, which can solve both fused Lasso and GraphNet regularized SVM with very little modification. We also demonstrate that the inner subproblems of the algorithm can be solved efficiently in analytic form by coupling the variable splitting strategy with a data augmentation scheme. Experiments on simulated data and resting state scans from a large schizophrenia dataset show that our proposed approach can identify predictive regions that are spatially contiguous in the 6D “connectome space," offering an additional layer of interpretability that could provide new insights about various disease processes.

Low-rank generalized linear regression for link prediction

Yun-Jhong Wu, Statistics, yjwu@umich.edu

As one of the fundamental problems, link prediction can be formulated as graph on estimation and treated as non- or semi-parametric regression on a product space. However, popular techniques such as kernel smoothing may fail due to unidentifiability of graphons. To overcome this difficulty, we develop a generalized linear model to estimate graphons by using low- rank approximation. This model can adapt to utilize the information on node and edge covariates. In practice, this method performs well for various simulation settings and real data in terms of both the accuracy of prediction and computational efficiency.

High Dimensional Covariance Matrix Estimation via the Barra Model

Yiwei Zhang, Statistics, evyzhang@umich.edu

The Barra model is one of the most popular risk models in financial industry for estimating the covariance matrix of financial assets, and the Barra one-step and two-step approaches are widely used to implement the estimation. In this paper, we first examine theoretical properties of the Barra model, which have somehow been ignored in the literature. In particular, we investigate the impact of the sample size (i.e., the number of trading days) and the number of financial assets on the performance of the Barra model.

We show that as the sample size increases, the Barra approach, unlike the sample covariance, is in fact not asymptotically consistent. This result is a little surprising and has never been reported. On the other hand, when the sample size is fixed and the number of financial assets increases, which is more realistic in practice, we show that the Barra approach outperforms the sample covariance. To further improve the estimation, we re-interpret the Barra model via the framework of the random effects model and propose an EM-like method to estimate the covariance. We show that under certain conditions, the new method is asymptotically consistent when the sample size increases, and when the sample size is fixed while the number of financial assets increases, the new method performs as well as the traditional Barra approach. Extensive simulation studies are used to support the theoretical results and compare the Barra approach, the new method and the sample covariance.

Poster Session III 3:00 PM – 4:00 PM

Coercive grain to grain registration for microscopy images

Yu-Hui Chen, Electrical Engineering and Computer Science, yuhuic@umich.edu

Image registration between different modality data has been widely investigated since it is an inevitable step to fuse the information from different sensors/scanners. However, without a complete image formation forward model, the interpolation during registration might introduce artificial noise and bias. In this work, we propose a coercive approach to register and segment the multimodality images which share similar sub-structure without interpolation. We demonstrate that our approach has significant better performance than the state-of-the-art registration and segmentation methods on microscopy images.

Regularized Block Toeplitz Covariance Matrix Estimation via Kronecker Product Expansions

Kristjan Greenewald, Electrical Engineering and Computer Science, greenewk@umich.edu

In this work we consider the estimation of spatio-temporal covariance matrices in the low sample non-Gaussian regime. We impose covariance structure in the form of a sum of Kronecker products decomposition (Tsiligkaridis et al. 2013, Greenewald et al. 2013) with diagonal correction (Greenewald et al.), which we refer to as DC-KronPCA, in the estimation of multiframe covariance matrices. This paper extends the approaches of (Tsiligkaridis et al.) in two directions. First, we modify the diagonally corrected method of (Greenewald et al.) to include a block Toeplitz constraint imposing temporal stationarity structure. Second, we improve the conditioning of the estimate in the very low sample regime by using Ledoit-Wolf type shrinkage regularization similar to (Chen, Hero et al. 2010). For improved robustness to heavy tailed distributions, we modify the KronPCA to incorporate robust shrinkage estimation (Chen, Hero et al. 2011). Results of numerical simulations establish benefits in terms of estimation MSE when compared to previous methods. Finally, we apply our methods to a real-world network spatio-temporal anomaly detection problem and achieve superior results.

Estimating Treatment Effects for Masked Trials Under Missing Not at Random

Shan Kang, Biostatistics, shankang@umich.edu

We focus on a missing not at random (MNAR) model for masked experiments in clinical trials. Missing data is an unavoidable problem in the most clinical trials. Most existing missing data approaches focus on missing at random (MAR) model. However, the MAR assumption is very questionable when the real causes of missing data are not well known, and cannot be tested from the current data. We propose a specific MNAR assumption which may be more plausible than MAR assumption for masked trials. We study models for categorical and continuous models under this assumption. Simulations are conducted to examine the finite sample performance and compare the results with methods that assume MAR. This idea is also applied to a clinical trial data example.

Optimization via Low-rank Approximation with Applications to Network Community Detection

Can Minh Le, Statistics canle@umich.edu

The community detection is an important problem in network analysis. Several methods have been proposed to solve the problem, including spectral clustering, modularity, and likelihood-based methods. One issue that many of such methods have to deal with is the optimization problem over a discrete set of labels. In this paper we introduce a general approach for solving the problem of maximizing a network criterion by projecting the set of labels onto a subspace spanned by leading eigenvectors of the network adjacency matrix. The main idea is that projection onto a low-dimensional space makes the feasible set of labels much smaller and the optimization problem much easier. By applying our method to maximize network likelihoods, we also provide insight into the connection between spectral clustering and likelihood-based methods. Simulations and application to real-world data show that our method performs well over a wide range of parameters.

Link Prediction Using Network Topology and Node Covariates

Bopeng Li & Ambuj Tewari, Statistics, bopengli@umich.edu

Link prediction is an important and open problem in network analysis. Many methods have been proposed for solving the problem, but most of them only take into account network topological covariates, such as the Jaccard Distance, while pay little attention to additional information on the nodes that is often available in practice. We develop a kernel-based learning method that uses both network topology and node covariates to do link prediction. We show how these covariates can be combined to achieve better prediction accuracy than using either type of covariates exclusively. Moreover, our method has the flexibility of learning the impact of covariates from different sources by combining multiple kernels instead of using a single one. We also propose an approach to address the class imbalance problem, which is common in real world networks, by optimizing a loss function with different weights on different classes. Experiment results on simulated and real network data show the efficacy of our method.

Network-Based Pathway Enrichment Analysis with Incomplete Network Information

Jing Ma, Statistics, mjing@umich.edu

Pathway enrichment analysis has become a key tool for biomedical researchers to gain insight in the underlying biology of differentially expressed genes, proteins and metabolites. It reduces complexity and provides a systems-level view of changes in cellular activity in response to treatments and/or progression of disease states. Methods that use pathway topology information have been shown to outperform simpler methods based on over-representation analysis. However, despite significant progress in understanding the association among members of biological pathways, and expansion of new knowledge data bases, such as KEGG, Reactome, BioCarta, etc., the existing network information may be incomplete/inaccurate, and are not condition-specific. We propose a constrained network estimation framework that combines network estimation based on cell- and condition-specific omics data with interaction information from existing data bases. The resulting pathway topology information is subsequently used to provide a framework for simultaneous testing of differences in mean expression

levels, as well as interaction mechanisms. We study the asymptotic properties of the proposed network estimator and the test for pathway enrichment, and investigate its small sample performance in simulated experiments and on a bladder cancer study involving metabolomics data.

Ensemble estimation of multivariate f-divergence

Kevin Moon, Electrical Engineering and Computer Science, krmoon@umich.edu

f-divergence estimation is an important problem in the fields of information theory, machine learning, and statistics. While several divergence estimators exist, relatively few of their convergence rates are known. We derive the MSE convergence rate for a density plug-in estimator. Then by applying the theory of optimally weighted ensemble estimation, we derive a divergence estimator with a convergence rate of $O\left(\frac{1}{T}\right)$ that is simple to implement and performs well in high dimensions. We validate our theoretical results with experiments.

Joint Semiparametric Time-to-Event Modeling of Cancer Onset and Diagnosis When Onset is Unobserved

John Rice & Alex Tsodikov, Biostatistics, jdrice@umich.edu

In cancer research, interest frequently centers on factors influencing onset of disease. However, in practice it is impossible to observe the time of onset precisely, making inference about this process difficult. To address this problem, we propose a joint model for the unobserved time to onset and time to diagnosis, with the two events linked by the baseline hazard. Covariates enter the model parametrically as linear combinations that multiply, respectively, the hazard for onset and the hazard for diagnosis conditional on onset. The baseline hazard is estimated nonparametrically using the EM algorithm, which allows for closed-form Breslow-type estimators at each iteration, drastically reducing computational time compared with maximizing the marginal likelihood directly. The parametric part of the model is estimated by maximizing the profile likelihood. We present simulation studies to illustrate the finite-sample properties of the method; its use in practice is demonstrated in the analysis of a prostate cancer data set.

Statistical Strategies for Constructing Health Risk Models with Multiple Pollutants and Their Interactions

Zhichao Sun, Yebin Tao, Shi L, Kelly K Ferguson, John D Meeker, Sung Kyun Park, Stuart A Batterman, & Bhramar Mukherjee, Biostatistics, zcs@umich.edu

Estimating the adverse health effects due to simultaneous exposure to multiple pollutants is an important topic to explore, and its challenges reside in, but are not limited to: identification of the most critical components of the pollutant mixture, examination of potential interaction effects, and attribution of health effects to individual pollutants in the presence of multicollinearity. In this study, we reviewed five methods available in the statistical literature and conducted a simulation study evaluating their performances. We also proposed a two-step strategy employing an initial screening by a tree-based

method followed by further dimension reduction/variable selection at the second step. From our investigation, there is no uniform dominance of one method across all simulation scenarios. Least absolute shrinkage and selection operator regression performs well for identifying important exposures, but will yield biased estimates and slightly larger model dimension given extensive collinearity and modest sample size. Bayesian model averaging and supervised principal component analysis are useful in variable selection under a strong exposure-response association. Substantial improvements on reducing model dimension and identifying important variables have been observed for the two-step modeling strategy when a large number of candidate variables exist, implying its potential under a multipollutant framework.

Systematic characterization of a wide spectrum of indels via repeat-aware hidden Markov models

Adrian Tan, Biostatistics, atks@umich.edu

Accurate and comprehensive detection of short insertions and deletions (indels) from shotgun sequence reads remains an elusive goal compared to single nucleotide variants (SNVs). For example, different indel calling methods show relatively low concordance (50---72%) compared to SNVs (83---93%) in recent evaluations from the 1000 Genomes Project. While many indels are located in uniquely mappable regions of genome with high sequence diversity (isolated indels), a large fraction (~50%) of indels are located in repetitive regions of genome, mostly in the form of short tandem repeats (STR), often with inexact repeat units. The spectrum of indels encompasses these two extreme forms, and detection of indels becomes progressively more difficult as the length of indels increases and as nearby sequence diversity decreases. To be an effective and robust indel caller, the calling algorithm must be aware of the inherent heterogeneous nature of indels.

We propose a set of pair hidden Markov models (Pair HMM) that covers the spectrum of indels by allowing us to explicitly model inexact repeats that often complicates the alignment of sequence data containing indels in repeat---rich regions of genome. Our approach first identifies the repeat unit in an inserted or deleted allele; using this candidate repeat unit, we then perform HMM alignments iteratively to determine appropriate 5'- and 3'- flanking sequences. With these three pieces of information in place, we construct a pair HMM that explicitly models both flanking sequences and the repeat unit, allowing for additional mismatches or indels within each repeat unit. This allows us to compute genotype likelihoods of indels found across the spectrum and facilitates the collection of additional information on how well the data fit to the model, which may potentially be useful for indel filtering. We evaluate our method by applying our model to deeply sequenced trios in the 1000 Genomes Project.

Robust nonparametric profile monitoring

Jingshen Wang & Changliang Zou, Statistics, jshwang@Umich.edu

Profile monitoring is a technique for checking the stability of the functional relationship between a response variable and one or more explanatory variables over time. General profile monitoring via nonparametric regression is particularly useful in practice due to its simplicity and flexibility. The existing monitoring methods suffer from a drawback in that they all assume the error distribution to be normal and accordingly local polynomial regression is used for nonparametric regression. However, the efficiency of least-squares (LS) based methods is adversely affected by outlying observations and heavy tailed distributions. To overcome this issue, we propose a robust monitoring EWMA procedure by incorporating with a rank-based test. Benefiting from certain favorable properties of such rank-based test, the proposed chart is robust from both the IC and OC ARLs' point of view, particularly when the process distribution is heavily tailed. An example with real data from the manufacturing industry shows that it performs well in application.

Multivariate ranking

Yingchuan Wang, Statistics, yingcw@umich.edu

Multivariate ranking has played an increasingly more crucial role in statistics. Most statistical measurements are multivariate by nature, and due to the development of innovative computer technologies, more and more large scale multivariate datasets are being analyzed. Ranking of data is fundamental to many statistical procedures. Because multivariate data ranking is not unique, many different ranking methods have been proposed over the years by using different data depth functions, but they all have pros and cons. For instance, the Mahalanobis depth works quite well when the underlying structure of the data is elliptical symmetric, but it is quite sensitive to outliers; the simplicial depth works well in low-dimensional cases, but the computational complexity of this method renders it impractical in high dimensions or large data size. We explore a new ranking method by transformation of the data to a common space. A very interesting fact is that for almost all kinds of data, the transformation leads to the same stationary distribution for the data ranks.

Measuring Influence in Social Networks through Information Diffusion Modeling

Donggeng Xia, Statistics, donggeng@umich.edu,

Data extracted from social network communication platforms, such as Twitter, records users' interactions over time. A key question is to determine who are the most influential members in such networks. A common influence measure discussed in the literature is based on a variant of the popular Page-Rank algorithm. In this talk, we propose a modification of such rank prestige algorithms by modeling the weights used to reflect users' activity, as opposed to users connectivity that is currently the case. The users activity is captured through interactions between users on the information they post, rebroadcast or comment on. These activities are modeled as multivariate interacting counting processes. We discuss how to estimate their parameters through maximum likelihood and establish their asymptotic properties. The proposed model is illustrated on simulated data as well as a data set extracted from discussions on Twitter.

Measuring Non-response Bias in Web Surveys: The Role of Health Status

Mengmeng Zhang, Ting Yan, Lindsay Ryan, & Jacqui Smith, Survey Methodology, zhanmeng@umich.edu

Web surveys have been included in many national longitudinal studies and panels as an additional option offered to respondents together with, or instead of, traditional modes (e.g. CATI or CAPI). Web surveys cost less and this mode generally requires a shorter data collection time compared to the traditional modes of data collection. Evidence is mixed regarding the presence and size of coverage error and non-response error in Web surveys. Typically, there is little information about non-respondents to allow for the measurement of non-response bias.

This paper takes advantage of the 2011 Health and Retirement Study (HRS) Internet Survey to examine factors affecting people’s decisions to respond to a Web survey. Respondents to the 2011 HRS Internet survey are randomly selected from respondents to the 2009 HRS internet survey and the 2010 HRS respondents who had Internet access. As a result, a lot of information is available on both respondents and non-respondents to the 2011 HRS Internet survey.

Making use of the rich auxiliary data, we first characterize non-respondents to the 2011 HRS Internet survey. Specifically, we examine whether health status is a factor that might influence respondents’ willingness to participate in the Web survey using logistic regression models controlling for age, gender, and socioeconomic status (SES). We hypothesize that respondents with lower self-rated health in previous waves were less likely to participate in the 2011 Internet survey. We then explore the role of health status in nonresponse bias in several key survey variables from the 2011 Internet survey for older adults. The final step is to predict people’s participation of the internet survey using logistic regression modeling based on auxiliary variables and key survey variables and compare the differences between the predicted results and the actual observations.

Multiple Imputation in Two-Stage Cluster Samples Using the Weighted Finite Population Bayesian Bootstrap

Hanzhi Zhou, Michael R. Elliott, & Trivellore E. Raghunathan, Survey Methodology, zhouhanz@umich.edu

Multiple imputation (MI) is a well-established method in dealing with item-level missing data and has become increasingly popular in the public health and social science investigations where data production is often based on complex sample surveys. However, existing software packages and procedures typically do not incorporate complex sample design features in the imputation process. We extend the “two-step MI” framework of Zhou et al. (2013) and utilize the weighted finite population Bayesian bootstrap to accommodate clustering effects in addition to the effect due to sample weights in two-stage unbalanced cluster samples. We propose two different procedures to simulate the unsampled part of population elements, both are able to produce draws from the posterior predictive distribution of the population that incorporate both clustering and weighting components of the sample design and for which missing data elements can then be imputed under an IID assumption. While the framework of fully parametric MI does not seem to provide a direct and robust technique to deal with sample weights in hierarchical imputation models, our method turns out to be a potential alternative that recovers most of information in the data generating mechanisms. We apply the different MI approaches to the analysis of passenger vehicle injury data from the National Automotive Safety System – Crash Detection System (NASS-CDS) survey.

A Shrinkage Estimator of Log Odds Ratio for Comparing Mobility Tables

Xiang Zhou, Statistics, xiangzh@umich.edu

In this paper, I present a shrinkage estimator of the log odds ratio for comparing mobility tables. Building on the James-Stein estimation rule and an empirical Bayes derivation, this estimator borrows information across multiple tables while placing no restrictions on the structure of association within tables. Numerical simulation shows that the shrinkage estimator outperforms the usual maximum likelihood estimator (MLE) in both the total squared error and the correlation with the true values. Moreover, the benefits of the shrinkage estimator relative to the MLE depend on the variation in the true log odds ratio and the variation in sample size among tables. To illustrate the effects of shrinkage, I contrast the shrinkage estimates with the usual estimates for the mobility data assembled by Hazelrigg and Garnier for 16 countries in the 1960s and 1970s. For mobility tables with more than two categories, the shrinkage estimates of log odds ratios can also be used to calculate summary measures of association that are based on aggregations of log odds ratios. Specifically, I construct an adjusted estimator of the Altham index, and, with a set of calibrated simulations, demonstrate its usefulness in both reducing the mean squared error and improving the correlation with the true values. Finally, using two real data sets, I show that in terms of gauging the overall degree of social fluidity, the adjusted estimates of the Altham index agree more closely with results from the Unidiff model than do direct estimates of the Altham index.