Workshop on BFF (Bayes, Fiducial and Frequentist) Paradigm in Data Integration, Machine Learning and Applications

9:00AM – 4:30PM, November 2, 2019

Room 1690 SPH I, School of Public Health, University of Michigan

Morning Scientific Program

9:00-10:30 BFF Approaches to Complex Biomedical Data Analysis

Chair: Peter Song, University of Michigan

9:00-9:30 Heping Zhang, Yale University School of Public Health

Title: It's the Interaction, Stupid

9:30-10:00 Annie Qu, University of Illinois at Urbana-Champaign

Title: Correlation Tensor Decomposition and Its Application in Spatial Imaging Data

10:00-10:30 Kevin He, University of Michigan School of Public Health

Title: Modeling time-varying effects in large-scale survival analysis

10:30-10:40 **Coffee/Tea Break**

10:40-12:10 BFF or Non-BFF Approaches to Data Integration

Chair: Jian Kang, University of Michigan

10:40-11:10 Jeremy M.G. Taylor*, Tian Gu, Bhramar Mukherjee, University of Michigan

Title: Empirical Bayes Approach to Integrate Summary-level Information from Multiple External Studies into the Current Study

11:10-11:40 Peisong Han, Jeremy Taylor, Bhramar Mukherjee University of Michigan

Title: Integrating Information from Existing Risk Calculators into Regression Model Fitting

11:40-12:10 Philip Boonstra* and Ryan P. Barbaro, University of Michigan

Title: Incorporating historical information with adaptive Bayesian updates

12:10-13:30 **Lunch Break**

Afternoon Scientific Program

13:30-14:30 **Non-asymptotic BBF Inference in Action**

Chair: Gongjun Xu, University of Michigan

13:30-14:00 Peng Wang*, University of Cincinnati, and Minge Xie, Rutgers University

Title: Repro Sampling Method for Joint Inference of Model Selection and Regression Coefficients in High Dimensional Linear Models

14:00-14:30 Chuanhai Liu, Purdue University

Title: Elucidating Foundations of Statistical Inference with the Cauchy Distribution

14:30-14:40 **Coffee/Tea Break**

14:40-16:30 **Rising Stars in BFF**

Chair: Jonathan Boss, University of Michigan

14:40-15:05 Emily Hector* and Peter Song, University of Michigan

Title: A Unifying Framework for Distributed and Integrated Inference with High-Dimensional Correlated Outcomes

15:05-15:30 Tian Gu*, Jeremy M.G. Taylor, Wenting Cheng and Bhramar Mukherjee

University of Michigan

Title: Synthetic Data Method to Incorporate External Information into the Current Study

15:30-15:55 Lan Luo* and Peter Song, University of Michigan

Title: Streaming Data Integration in Real-time Regression Analysis

15:55-16:20 Yiwang Zhou* and Peter Song, University of Michigan

Title: Synergistic Self-learning of Individualized Dietary Supplement Rules from Multiple Health Benefit Outcomes

16:20-16:30 Closing Remarks

Note: Name accompanied with * is the name of speaker.

Abstracts (Alphabetical Order by Speaker)

Philip S Boonstra

In this talk, I will discuss Bayesian approaches for incorporating information from a historical model into a current analysis when the historical model includes only a subset of covariates currently of interest. The statistical challenge is two-fold. First, the parameters in the nested historical model are not generally equal to their counterparts in the larger current model, neither in value nor interpretation. Second, because the historical information will not be equally informative for all parameters in the current analysis, additional regularization may be required beyond that provided by the historical information. I propose several novel extensions of the power prior that adaptively combine a prior based upon the historical information with a variance- reducing prior that shrinks parameter values toward zero. The ideas are directly motivated by my work building mortality risk prediction models for pediatric patients receiving extracorporeal membrane oxygenation, or ECMO. We have developed a model on a registry-based cohort of ECMO patients and now seek to expand this model with additional biometric measurements, not available in the registry, collected on a small auxiliary cohort. My adaptive priors are able to use the information in the original model and identify novel mortality risk factors. I support this with a simulation study, which demonstrates the potential for efficiency gains in estimation under a variety of scenarios.

Peisong Han

Consider the setting where (i) individual-level data are collected to build a regression model for the association between observing an event of interest and certain covariates, and (ii) some risk calculators predicting the risk of the event using less detailed covariates are available, possibly as black boxes with little information available about how they were built. We propose a general empirical-likelihood-based framework to integrate the rich auxiliary information contained in the calculators into fitting the regression model in order to improve the efficiency for the estimation of regression parameters. Two methods are developed, one using working models to extract the calculator information and one making a direct use of calculator predictions without working models. Both theoretical and numerical investigations show that the calculator information can help substantially reduce the variance of regression parameter estimation. As an application, we study the dependence of the risk of high grade prostate cancer on both conventional risk factors and newly identified biomarkers by integrating information from the Prostate Biopsy Collaborative Group (PBCG) risk calculator, which was built based on conventional risk factors alone. This is joint work with Jeremy Taylor and Bhramar Mukherjee.

Kevin He

National disease registries have produced a vast amount of data. Many existing statistical methods that perform well for moderate sample sizes and small-dimensional data do not scale to such large-scale data, leading to a demand for statistical techniques that enable full utilization of these rich sources of information. For example, the time-varying effects model is a flexible and powerful tool for modeling the dynamic changes of covariate effects. However, in survival analysis, its computational burden increases quickly as the number of sample sizes or predictors grows. Analyses with a massive sample size and large number of predictors defy any existing statistical methods and software. In view of these difficulties, we propose a Minorization-Maximization-based Block-Coordinate Ascent method for estimating the time-varying effects. Leveraging the block structure formed by the basis expansions, the proposed procedure iteratively updates the optimal block-wise direction along which the approximate increase in the log-partial likelihood is maximized. The resulting estimates ensure the ascent property and serve as refinements of the previous step. The performance of the proposed method is examined by simulations and applications to the analysis of national kidney transplant data and cancer death data from the U.S. SEER cancer registry.

Emily Hector

This talk is motivated by a study of infant memory that involves a regression analysis of electroencephalography (EEG) neuroimaging data with high-dimensional correlated responses with multi-level nested correlations. We propose a general class of distributed estimators that can be implemented in a fully parallelized computational scheme for estimation and inference with high-dimensional correlated outcomes. Modelling, computational and theoretical challenges related to high-dimensional correlated outcomes are overcome by first fitting many local models within subsets of the data and then combining local results in a statistically efficient way. Our approach to distributed estimation and inference is formulated using Hansen's generalized method of moments, and is statistically efficient. We provide rigorous theoretical justifications for the use of distributed estimators with correlated outcomes by studying the asymptotic behaviour of the combined estimator with fixed and diverging number of data divisions. We develop an R package for ease of implementation

Tian Gu

In the big data era, incorporating external summary-level information into current study has attracted significant interest to improve the estimation efficiency. We consider the situation where there is a known regression model that can be used to predict an outcome of interest from a set of commonly available predictors. An internal modest-sized dataset is available containing individual level data for the variables in the known model as well as a new variable. The challenge is to build an improved prediction model that includes the new variable, using both the internal individual level data and information obtained from the external known model. We propose a synthetic data approach, which consists of using the known model to create synthetic data observations with missing values of the new variable, and then appending them to the internal data to create a combined dataset incorporating the external information from the known model. To estimate the parameters of the improved model, this combined dataset is analyzed using methods that can handle missing data (e.g. multiple imputation). A theoretical justification of the method is provided, and it is evaluated in simulation studies. The method is applied to improve models for the risk of prostate cancer. The method's broad applicability makes it appealing for use across diverse scenarios.

Chuanhai Liu

Statistical inference as a formal method in scientific investigations to covert experience to knowledge has proven to be elusively difficult. While frequentist and Bayesian methodologies have been accepted in the contemporary era as two dominant schools of thought, it has been a good part of the last hundred years to see growing interests in development of more sound methods, both philosophically, in terms of scientific meaning of inference, and mathematically, in terms of exactness and efficiency. These include Fisher's fiducial, Dempster-Shafe theory of belief functions, generalized fiducial, Confidence Distributions, and the most recently proposed inferential framework, called Inferential Models. Since it is notoriously challenging to make exact and efficient inference about the Cauchy distribution, this talk takes it as an example to elucidate different schools of thought on statistical inference. It is shown that the standard approach of Inferential Models produces exact and efficient prior-free probabilistic inference on the location and scale parameters of the Cauchy distribution, whereas all other existing methods suffer from various difficulties.

Lan Luo

Abstract: Streaming data integration refers to integrating data in real-time to provide up-to-date information. The need for streaming data integration has emerged due to the increase in information sources where large amount of observations are collected sequentially and perpetually over time, including national disease registry, mobile health and disease surveillance, among others. This work primarily concerns the development of a fast real-time statistical methodology for regression analysis, with an attempt to integrate streaming data in form of summary statistic, rather than subject-level data. Termed as renewable estimation, this method helps greatly overcome data sharing barrier, reduce data storage cost and improve computing speed. More importantly, the difference between our renewable estimator and the oracle one obtained by using all subject-level data vanishes as the total sample size increases, and such estimation consistency is more advantageous comparing to divide-and-conquer scheme. The proposed algorithms for streaming data integration will be demonstrated in generalized linear models (GLM) for cross-sectional data. Numerical examples from both simulation experiments and a real-world data analysis will be used to illustrate the performance of this method.

Annie Qu

Most of existing statistical models in imaging analysis only focus on the first moment information of imaging pixels, while the important pixel-wise correlation structure is usually ignored. In this talk, motivated by the multimodal optical imaging data in a breast cancer study, we propose a new tensor learning approach to analyze spatial-correlated imaging data. Specifically, we construct a higher-order correlation tensor which effectively preserves the spatial information and captures the pixel-wise correlation structure. In addition, we propose a new semi-symmetric tensor decomposition method to model spatial correlations, which enables us to identify spatial structures associated with disease, and thus improves the diagnostic power. We also establish the theoretical properties for recovering the true spatial correlation structure, and develop scalable computational algorithm. We illustrate the performance of the proposed method in both simulation studies and the application to multi-photon breast cancer imaging data. The numerical results indicate that the proposed method outperforms other competing methods including the Convolutional Neural Network (CNN), especially when the sample size of imaging data is limited. This is joint work with Yujia Deng and Xiwei Tang.

Jeremy M.G. Taylor

We consider the situation in which there are K external studies, each of which developed a prediction model for the same outcome. Each of the external studies may use a slightly different set of covariates. The parameters of the external models are known, but the external data is not available. The goal is to develop a prediction model that uses all the possible covariates, using data from an internal study. The parameters of this model are the quantities of interest. We propose a meta-analysis framework, in which the parameters of interest may differ between the internal and external populations, but are assumed to be drawn from a common distribution. We use an empirical Bayes estimation approach, which uses the estimates from the different external models. The approach first separately incorporates the different summary information from each external study into the internal study, and then takes a weighted average of the resulting estimates to give a final overall estimate of the parameters of interest. This estimate is more efficient than the simple analysis of the internal data. The approach also gives an empirical best linear unbiased estimator for each of the internal and external populations.

Peng Wang

This paper proposes a new and effective simulation-based approach, called Repro Sampling method, to conduct statistical inference in high dimensional linear models. The Repro method creates and studies the performance of artificial samples (referred to as Repro samples) that are generated by mimicking the sampling mechanism that generated the true observed sample. By doing so, this method provides a new way to quantify model and parameter uncertainty and provide confidence sets with guaranteed coverage rates on a wide range of problems. A general theoretical framework and an effective Monte-Carlo algorithm, with supporting theories, are developed for high dimensional linear models. This method is used to joint create confidence sets of selected models and model coefficients, with both exact and asymptotic inferences are included. It also provides theoretical development to support the computational efficiency. Furthermore, this development allows us to handle inference problems involving covariates that are perfectly correlated. A new and intuitive graphical tool to present uncertainties in model selection and regression parameter estimation is also developed. We provide numerical studies to demonstrate the utility of the proposed method in a range of problems. Numerical comparisons suggest that the method is far better (in terms of improved coverage rates and significantly reduced sizes of confidence sets) than the approaches that are currently used in the literature. The development provides a simple and effective solution for the difficult post-selection inference problems.

Heping Zhang

The vast majority of statistical methods, theory, and applications are for or based on additive models, and often linear models. Those models have led to reasonable and easily interpretable models. However, in practice, more likely than not, the underlying data structures are not additive, and the only difference is to what degree the additive models provide a sufficiently good fit to the data. For example, in the analysis of data from genomewide association studies, the failure to identify genes with major effects on complex diseases is very likely due to our inability to consider and identify gene-gene and/or gene-environment interactions. In this talk, I will present two approaches to detecting interactions and demonstrate their potential with numerical examples.

Yiwang Zhou

Deriving and validating individualized dietary supplement rules (IDSRs) based on attributable effects of nutrients on health is a notoriously hard problem in precision nutritional sciences, where multiple benefit outcomes are often available in analyses. Two major challenges arise in the development of IDSRs in the presence of multiple variables pertinent to health benefit, including (i) individual outcomes are of different clinical relevance to the overall underlying health benefit, and (ii) outcomes are measured separately from different subgroups of subjects, leading to different sample sizes and missing data patterns across outcomes. This paper is motivated from a randomized clinical trial that aims to assess the effect of calcium supplementation for pregnant women on the reduction of infant's in utero exposure to lead. We propose to integrate different types of blood lead concentration measurements of varying clinical relevance and sample sizes in order to create a larger training dataset in the training of a self-learning scheme for IDSRs. We develop a new support vector machine (SVM) that allows us to synergize heterogeneous training data sources in a weighted self-learning paradigm. As a significant extension to the existing outcome weighted learning (OWL), our proposed synergistic self-learning (SS-learning) incorporates both multiple outcomes and data missingness in the optimization, resulting in an optimal solution of individual rules for dietary supplement with respect to a composite health benefit. We establish the algorithmic convergence of the proposed SS-learning, and illustrate the performance of this methodology through both simulation studies and real data analysis of the motivating calcium supplementation clinical trial.