1. POSTERS: VARIABLE SUBSET SELECTION

1a. TOTAL VARIATION DENOISING OF COEFFICIENT FUNCTIONALS IN THE ADDITIVE COMPLEMENTARY LOG-LOG SURVIVAL MODEL

Hao Sun*, University of Rochester
Brent A. Johnson, University of Rochester

In the analysis of censored failure time data, temporal covariate effects often indicate meaningful heterogeneity in covariate-response associations and the analysis assuming time invariant covariate effects may mislead such underlying relationship. This article describes a variant of temporal covariate effects model (Peng and Huang 2007) which is an adaptive smoothing-free approach to directly study the time varying effect of covariates to survival function for survival data. We present a simple penalization-based framework to improve the operating characteristics of this novel estimator. The resulting estimate is consistent, more efficient and less complex. Extensive simulation studies are conducted to verify the finite sample performance and a HIV retention application is used to illustrate the practical utility of the proposed method.

email: hao_sun@urmc.rochester.edu

1b. BAYESIAN VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA WITH ORDINAL RESPONSES

Yiran Zhang*, The Ohio State University
Kellie J. Archer, The Ohio State University

Health outcome and disease status measurements frequently appear on an ordinal scale, that is, the outcome has inherent ordering. Many previous studies have shown associations between gene expression and disease status. Identification of important genes may be useful for developing novel diagnostic and prognostic tools to predict or classify stage of disease. Gene expression data is usually high-dimensional, meaning that the number of genes is much greater than the sample size or number of patients. In this poster, we will describe some existing frequentist methods for high-dimensional data with an ordinal response. Following Tibshirani (1996) who described the LASSO estimate as the Bayesian posterior mode when the regression parameters have independent Laplace priors, we propose a new approach for high-dimensional data with an ordinal response that is rooted in the Bayesian paradigm. We show that our proposed Bayesian approach outperforms the existing frequentist methods through simulation studies. We then compare the performance of frequentist and Bayesian approaches using real data applications.

email: zhang.4830@osu.edu

1c. SELECTING APPROPRIATE PROBABILISTIC MODELS FOR MICROBIOME DATA ANALYSIS

Hani Aldirawi*, University of Illinois at Chicago
Jie Yang, University of Illinois at Chicago
Ahmed A. Metwally, Stanford University

The human microbiome plays an important role in human disease and health. There is a strong association between microbiome and host-immune response and multiple diseases. In this presentation, we first briefly review the background and significance of the microbiome, the technologies used for collecting microbiome data, and some public resources for downloading microbiome data, and the probabilistic models used in the literature for read counts from a specific feature. We then introduce a general procedure and test for choosing appropriate probabilistic models for microbiome data analysis, including normal, Poisson, negative binomial, half-normal, beta binomial, exponential, beta negative binomial, zero-inflated models, and Hurdel models.

email: haldir2@uic.edu

1d. ON THE MUST-BE OF VARIABLE SELECTION IN BIOMEDICAL RESEARCH

Bokai Wang*, University of Rochester
Changyong Feng, University of Rochester

A variable selection strategy called the Univariate Analysis Screening (UAS) is commonly implemented in top biomedical journals. In fact, it is called the marginal regression in statistics and also recommended as a way of finding the suite of candidate models under certain conditions in some top statistical journals. It's easy to show that this widely used variable selection procedure is problematic. Therefore, we propose a method called the Multi-splitting Backward Elimination (Must-Be), which can handle the variable selection problems in low-dimensional linear, logistic, Cox regression models with highly correlated predictors, and for moderate sample size. In addition, the Must-Be method is a generic approach, i.e., it could be extended to solve variable selection problems whenever asymptotic valid p-values are available. We compare our Must-Be procedure with a few academic/industrial recognized high-dimensional variable selection methods using extensive simulation studies. Our new procedure shows equivalent or better performance regarding minimizing the false negative rate and the false positive rate simultaneously under certain scenarios.

email: wang_bokai@hotmail.com

1e. AN ESTIMATION OF AVERAGE TREATMENT EFFECT USING ADAPTIVE LASSO AND DOUBLY ROBUST ESTIMATOR

Wataru Hongo*, Tokyo University of Science
Shuji Ando, Tokyo University of Science
Jun Tsuchida, Tokyo University of Science
Takashi Sozu, Tokyo University of Science

In observational studies, propensity score methods are often used to obtain an unbiased estimate of the average treatment effect (ATE). Usually, all confounders are selected as covariates for a propensity score model. Recent researches have revealed that the inclusion of all confounders and variables associated with the response variable in the propensity score model improved the bias and precision of ATE estimation. More recently, outcome-adaptive lasso has been proposed for the selection of appropriate covariates to be included in the propensity score model. However, when the estimated propensity score concentrates around 0 or 1, the bias in estimation of ATE increases, especially when the inverse probability of weighted methods is used. This study proposes a new method for estimating ATE, using both the outcome-adaptive lasso for constructing a propensity score model and the usual adaptive lasso for constructing a response model. We compared three methods:

(1) the proposed method; (2) the outcome-adaptive lasso method; and (3) the usual adaptive lasso method. We found that the bias of the proposed method was always lower than that of the other two methods.

email: 4418522@ed.tus.ac.jp

2. POSTERS: SURVIVAL ANALYSIS/COMPETING RISKS

2a. AUGMENTED DOUBLE INVERSE-WEIGHTED ESTIMATION OF DIFFERENCE IN RESTRICTED MEAN LIFETIMES USING OBSERVATIONAL DATA SUBJECT TO DEPENDENT CENSORING

Qixing Liang*, University of Michigan
Min Zhang, University of Michigan

Restricted mean lifetime is often of interest in medical studies. For example, cardiac surgeons are interested in comparing lifetimes between two types of ventricular assist device (VAD). However, the comparison is often complicated because types of VAD implants are not randomized and, more challengingly, it is not uncommon that patients receive heart transplant after the VAD implant, which may not be balanced between two study groups. We propose methods to estimate the treatment-specific difference in potential restricted mean lifetimes had no patients received heart transplant where heart transplant is treated as a dependent censoring for the potential lifetime. Specifically, we first derive an estimator that combines inverse probability of treatment weighting and inverse probability of censoring weighting to account for imbalance in baseline characteristics and receipt of heart transplant, respectively. Then we study various augmentation methods to improve the efficiency of estimation and compare their relative performances. Large-sample properties of the proposed methods are derived and simulation studies are conducted to assess the finite-sample performance.

email: liangqx@umich.edu

2b. SINGLE-INDEX MODELS WITH TRANSFORMATION MODELS FOR OPTIMAL TREATMENT REGIMES

Jin Wang*, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill

There is a growing interest in precision medicine, where a potentially censored survival time is often the most important outcome of interest. To discover optimal treatment regimens for such an outcome, we propose a semiparametric proportional hazards model by incorporating the interaction between treatment and a single index of covariates through an unknown monotone link function. We further extend this model to general forms of transformations to allow for various forms of the hazard functions. We propose a variable selection procedure for the single index term to accommodate high dimensional covariates and to select the relevant coefficients. We show that the estimators are consistent and asymptotically normal with a covariance matrix that attains the semiparametric efficiency bound, and the variable selection has the oracle property.

email: jinjin@live.unc.edu

2c. UNIVARIATE GRADIENT STATISTIC FOR MARGINAL CURE RATE MODEL WITH HIGH-DIMENSIONAL COVARIATES

Jennifer L. Delzeit*, Kansas State University
Jianfeng Chen, Kansas State University
Wei-Wen Hsu, Kansas State University
David Todem, Michigan State University
KyungMann Kim, University of Wisconsin, Madison

Cure rate models have been widely used in the literature for analyzing the lifetime data of long-term survivors. Owing to the advancement of genomics, it is of interest to identify genes that are highly associated with the survival outcome under the cure rate model. This identification procedure will involve variable selection for high-dimensional covariates. However, the cure rate model requires modeling of the cure fraction, which leads to a more complicated variable selection process. In this paper, we propose a gradient-statistic-based variable selection method under the marginal cure rate model. This marginal model can produce interpretable covariate effects on the overall survival response by relating the marginal mean hazard rate to high-dimensional covariates without specification of the cure fraction. A univariate gradient statistic is then used iteratively to determine significant covariates. Coupled with the use of a False Discovery Rate approach, the top-ranked list of covariates can be obtained. The results of the proposed method are illustrated by extensive simulations and an application to TCGA breast cancer dataset that contains over 400,000 microarrays.

email: jenny7@ksu.edu

2d. INTEGRATIVE SURVIVAL ANALYSIS WITH UNCERTAIN EVENT TIMES IN APPLICATION TO A SUICIDE RISK STUDY

Wenjie Wang*, University of Connecticut
Robert Aseltine, University of Connecticut
Kun Chen, University of Connecticut
Jun Yan, University of Connecticut

The concept of integrating data from disparate sources to accelerate scientific discovery has generated tremendous excitement in many fields. The potential benefits from data integration, however, may be compromised by the uncertainty due to imperfect record linkage. Motivated by a suicide risk study, we propose an approach for analyzing survival data with uncertain event times arising from data integration. We develop an integrative Cox proportional hazards regression, where the uncertainty is modeled probabilistically. The estimation procedure combines the ideas of profile likelihood and the expectation conditional maximization algorithm (ECM). Simulation studies demonstrate that under realistic settings of imperfect data linkage, the proposed method outperforms several competing approaches including multiple imputation. A marginal screening analysis using the proposed integrative Cox model is performed to identify risk factors associated with death following suicide-related hospitalization in Connecticut. The identified diagnostics codes are consistent with existing literature and provide several new insights on suicide risk prediction and prevention.

email: wenjie.2.wang@uconn.edu

2e. NONPARAMETRIC ESTIMATION OF THE JOINT DISTRIBUTION OF A SURVIVAL TIME AND MARK VARIABLE IN THE PRESENCE OF DEPENDENT CENSORING

Busola Sanusi*, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill

Joint models of survival time and mark variable are often times of interest in clinical trials. The problem of an induced correlation between the mark variable with the survival time is compounded when censoring mechanism is not independent. Inverse probability censoring weighted estimator approach has been used extensively to correct for dependent censoring. In this framework of joint modeling, previous literature did not account for information resulting from a censoring process and as a result, the estimated joint distribution function may yield biased results when survival times are dependent on censoring times. We consider nonparametrically estimating the joint distribution of a survival time and mark variable, where the survival time is subject to right censoring and the mark variable is only observed when the survival time is not censored. The possibility of dependent censoring is allowed for using inverse probability of censoring weights. Large sample properties of the proposed estimator will be presented. Finite sample behavior will be investigated via simulation study, and the proposed estimator will be illustrated using data from a recent HIV vaccine trial.

email: sanus1bo@email.unc.edu

2f. GROUP VARIABLE SCREENING FOR CLUSTERED MULTIVARIATE SURVIVAL DATA

Natasha A. Sahr*, St. Jude Children's Research Hospital
Kwang Woo Ahn, Medical College of Wisconsin
Soyoung Kim, Medical College of Wisconsin

The penalized regression methods for variable selection can achieve consistency for clustered survival data when the number of parameters is smaller than the number of clusters. However, penalized regression methods may be computationally expensive and unstable when the number of parameters is much larger than the number of clusters. In this setting, one can use a two-step procedure: i) screen group variables to make the number of covariates smaller than the number of clusters; and ii) use existing variable selection procedures to select significant group variables and within-group variables. We propose to extend the sure joint group screening method of Ahn et al. (2018) to the clustered multivariate survival setting using the marginal multivariate proportional hazards model. In this method, we address complicated data structures by allowing for screening groups across failure types, groups within failure types, and individual variables. Compared with existing methods modified for clustered multivariate survival data, the sure joint group screening method provides the optimal performance in simulation while maintaining both ascent and sure screening theoretical properties.

email: nsahrsimonis@gmail.com

2g. SEMIPARAMETRIC REGRESSION ON CUMULATIVE INCIDENCE FUNCTION WITH INTERVAL-CENSORED COMPETING RISKS DATA AND MISSING EVENT TYPE

Jun Park*, Indiana University
Giorgos Bakoyannis, Indiana University
Ying Zhang, Indiana University
Constantin T. Yiannoutsos, Indiana University

Cohort studies and clinical trials with time-to-event outcomes involve competing risk data. Two common issues in such studies are i) interval censoring that the exact event time is not precisely observed but is only known to lie between two clinic visits, and ii) partially observed event types. We address these two issues in the framework of semiparametric regression analysis of the cumulative incidence function. We impose a missing at random assumption conditionally on both the covariates of interest and auxiliary variables and propose an augmented inverse probability weighted sieve maximum likelihood estimator. Using empirical process theory, we show that the regression parameter estimator is root-n consistent, asymptotically normal and doubly-robust which means that it is consistent even if either the model for the probability of missingness or the model for the probability of the event type is misspecified. Simulation studies demonstrate good performance of the proposed method even under a large amount of missing event types. The method is illustrated using HIV data with a significant portion of missing event types from a cohort study in sub-Saharan Africa.

email: jp84@iu.edu

2h. LATENT CLASS REGRESSION MODELING OF COMPETING RISKS DATA

Teng Fei*, Emory Rollins School of Public Health
John Hanfelt, Emory Rollins School of Public Health and Emory Alzheimer's Disease Research Center
Limin Peng, Emory Rollins School of Public Health

We consider incorporating latent class analysis (LCA) with survival analysis to study the heterogeneity of a certain disease. The data used for LCA, however, may not show strongly distinct clustering patterns, such that hard labels generated from the LCA algorithms ignore the underlying ambiguities. In this project, we propose an estimation method for a joint model of latent classes and time to competing-risks events. Our proposal utilizes soft labels of membership probabilities for latent classes and enables valid assessment of the impact of latent classes on competing risks endpoints. We establish the asymptotic properties of the proposed estimator, including uniform consistency and weak convergence to a Gaussian process. Inferences for the time-dependent latent class effects are developed. Simulations and an application to a mild cognitive impairment (MCI) dataset demonstrate the practical utility of the proposed method.

email: tfei@emory.edu

2i. THE USE OF REPEATED MEASUREMENTS FOR DYNAMIC CARDIOVASCULAR DISEASE PREDICTION: THE APPLICATION OF JOINT MODEL IN THE LIFETIME RISK POOLING PROJECT

Yu Deng*, Northwestern University
Yizhen Zhong, Northwestern University
Abel Kho, Northwestern University
Lei Liu, Washington University School of Medicine
Norrina Allen, Northwestern University
John Wilkins, Northwestern University
Kiang Liu, Northwestern University
Donald Lloyd-Jones, Northwestern University
Lihui Zhao, Northwestern University

Multiple algorithms have been developed to predict the risk of cardiovascular disease (CVD), e.g., the pooled cohort equation. Those algorithms estimated the risk of CVD based on risk factors, such as sex, systolic blood pressure (SBP). However, most of them focused on values measured at a single time point. Repeated measurements provide more information and are likely to improve the algorithm performance. In this study, we derived a dynamic risk prediction model for CVD outcome by joint modeling the CVD event time and multiple longitudinal risk factors. The R package "Jmbayes" was used to fit the joint model. When using longitudinal data, irregular visits lead to missing data that cannot be handled well by the "Jmbayes". To overcome this challenge, we proposed a multiple imputation approach combined with joint model. We used data from the CVD Lifetime Risk Pooling Project. Our results suggest that joint models are promising in predicting CVD outcome. They had equivalent AUC scores compared to the pooled cohort equation when applied on the whole study population, and superior performance when applied on subgroup who have high variations in SBP.

email: yudeng2015@u.northwestern.edu

3. POSTERS: MACHINE LEARNING

3a. LEARNING IMAGE WITH GAUSSIAN PROCESS REGRESSION AND APPLICATION TO CLASSIFICATION

Tahmidul Islam*, University of South Carolina

Estimating distribution of pixel intensities of images is a challenging due to high dimensionality. An image is considered as a function over a 2D plane. Gaussian process (GP) regression is applied to estimate this function. The posterior predictive distribution obtained by regressing on the same spatial locations is then considered as the estimated distribution. For n images with (p x p) resolution, the GP regression has O(n3p6) computational complexity which makes implementation infeasible. We develop an ANOVA style representation, producing a computationally friendly and tractable expression requiring only O(p6) computation. We demonstrate the application of this image learning to classification using the Bayes classifier. We relax the conditional independence assumption of naive Bayes classifier (NBC) and use the predictive posterior distribution to develop a more general classifier. Our method outperforms the classification accuracy of NBC (52.40%) with 86.58% accuracy when applied to handwritten digit image dataset (MNIST). This technique can handle images with missing pixels and unequal resolution between training and test images.

email: islamt@email.sc.edu

3b. THE MODELS UNDERLYING WORD2VEC, A NATURAL LANGUAGE PROCESSING ALGORITHM, AND THEIR RELATIONSHIP TO TRADITIONAL STATISTICAL MULTIVARIATE METHODS

Brian L. Egleston*, Fox Chase Cancer Center, Temple University Health System
Tian Bai, Temple University
Slobodan Vucetic, Temple University

Natural language processing and artificial intelligence algorithms are increasingly being developed in the informatics fields. Many of these methods have not been well described within the traditional statistics literature. We present a description of a popular method, Word2Vec, and some of its similarities to cluster analysis and latent variable modeling. Reframing methods using more conventional probability theory notation may facilitate the incorporation of statistical rigor into their development.

email: Brian.Egleston@fccc.edu

3c. A TWO-STEP CLUSTERING ALGORITHM FOR CLUSTERING DATA WITH MIXED VARIABLE TYPES

Shu Wang*, University of Pittsburgh
Jonathan G. Yabes, University of Pittsburgh
Chung-Chou H. Chang, University of Pittsburgh

Clustering is an important technique in pattern discovering in various research areas. However, existing algorithms that can handle mixed types of data are limited. In addition, these methods usually have some drawbacks and will fail in some scenarios. In practice we almost always encounter data consist of mixed types of variables. One example is the recently popular EHR data. In this paper, we proposed a two-step clustering algorithm (2SCA) for mixed types of data. First step is pre-processing step which could filter out variables that have little contribution. Second step is final clustering step which applies a proposed dissimilarity measure on selected variables. We also defined 3 data structure types based on the information data can provide and demonstrated that our algorithm could always have good performances and provide variable importance information under all data structure types.

email: shw97@pitt.edu

3d. PseudoNet: RECONSTRUCTING PSEUDO-TIME IN SINGLE-CELL RNA-SEQ DATA USING NEURAL NETWORKS

Justin Lakkis*, University of Pennsylvania
Chenyi Xue, Columbia University
Huize Pan, Columbia University
Sarah B. Trignano, Columbia University
Hanrui Zhang, Columbia University
Nancy Zhang, University of Pennsylvania
Muredach Reilly, Columbia University

Gang Hu, Nankai University
Mingyao Li, University of Pennsylvania

Single-cell gene expression profiling can be used to quantify transcriptional dynamics in temporal processes using computational methods to label each cell with a "pseudo-time." Popular pseudotime reconstruction methods scale poorly as the number of cells increases. We present PseudoNet, a supervised machine learning approach for pseudotime reconstruction to model a non-branching temporal cell process. A neural network is trained to predict the probability that a cell is near a pre-specified end of the temporal process based on its gene expression. The predicted probability is used as a proxy for pseudo-time standardized to the [0, 1] scale, which allows for identification of genes that are differentially expressed over pseudo-time. We show that our method requires less computation time and classifies cells with greater accuracy than other popular pseudo-time reconstruction methods. We apply this method to a single-cell RNA-seq dataset consisting of monocyte gene expression data for 23 patients and show that PseudoNet identifies genes differentially expressed between healthy people and those with cardiometabolic disease risk factors.

email: jlakkis@pennmedicine.upenn.edu

3e. PATTERNS OF COMORBIDITY PRECEDING DEMENTIA DIAGNOSIS: FINDINGS FROM THE ATHEROSCLEROSIS RISK IN COMMUNITIES (ARIC) STUDY COHORT

Arkopal Choudhury*, University of North Carolina, Chapel Hill
Anna M. Kucharska-Newton, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

We propose to use the dementia classification available at Atherosclerosis Risk in Communities (ARIC) Visit 5, developed as part of the ARIC Neurocognitive Study (NCS), in combination with the linked ARIC CMS Medicare data, to examine comorbidities and patterns of healthcare use that precede the development of cognitive impairment (MCI and dementia). To answer our study questions, we propose to use machine learning methods, which will allow us to address the complexity of the comorbidity and health care use and their change over time to uncover patterns that may not have been apparent with traditional statistical approaches. We describe evidence of the clinical manifestation of cognitive impairment prior to the ARIC dementia and MCI classification. We apply text mining protocols to identify patterns of comorbidities associated with care preceding ARIC dementia and MCI classification, using the ICD-9 codes during hospitalization, to help characterize the patients affected by the disease.

email: arkopal@live.unc.edu

4. POSTERS: PERSONALIZED MEDICINE

4a. A UTILITY APPROACH TO INDIVIDUALIZED OPTIMAL DOSE SELECTION USING BIOMARKERS

Pin Li*, University of Michigan
Matthew Schipper, University of Michigan
Jeremy Taylor, University of Michigan

In Oncology, increasing the treatment dose generally increases both efficacy and toxicity. With continuous dose, the goal is to analyze an existing dataset to estimate an optimal dose for each (future) patient based on their clinical features and biomarkers. In this paper, we propose an optimal individualized dose finding rule by maximizing utility functions for each patient and satisfying the average toxicity tolerance. This approach maximizes overall efficacy at a prespecified constraint on overall toxicity. We model the outcomes using logistic regression with dose, biomarkers and their interactions. To incorporate the large number of biomarkers and their interactions with dose, we employ the LASSO with linear constraints on the dose related coefficients to constrain the dose effect to be non-negative. Simulation studies show that this approach can improve efficacy without increasing toxicity relative to fixed dosing. Constraining each patient estimated dose-efficacy and dose-toxicity curves to be non-decreasing improved performance relative to standard LASSO. The proposed methods are illustrated using a dataset of patients with lung cancer treated with radiation therapy.

email: pinli@umich.edu

4b. A COMPARISON AND ASSESSMENT OF RECENTLY DEVELOPED TREE-BASED METHODS FOR SUBGROUP IDENTIFICATION

Xinjun Wang*, University of Pittsburgh
Ying Ding, University of Pittsburgh

With rapid advances in understanding of the human diseases, the paradigm of the medicine shifts from one-fits-all to targeted therapies. Subgroup analysis and identification becomes a critical topic in clinical researches. The existence of treatment effect heterogeneity makes it important to identify subgroup of patients with enhanced efficacy for the treatment to target, or to decide the optimal treatment rule for any future patients based on their characteristics and treatment response histories. Many data-driven methods for subgroup analysis have been developed, most of which utilize machine learning techniques such as lasso and decision trees. Since decision trees are known to have advantages such as performing variable selection and providing easy interpretation, several tree-based methods, such as Virtual Twins, Interaction Trees and SIDES, are available for identification of subgroups with differential treatment effect. In this manuscript, we summarize the similarities and differences among five recent developed tree-based methods, and compare and assess the performance of those methods via extensive simulations that could represent real situations.

email: XIW119@pitt.edu

4c. A SIMULTANEOUS INFERENCE PROCEDURE TO IDENTIFY SUBGROUPS IN TARGETED THERAPY DEVELOPMENT WITH TIME-TO-EVENT OUTCOMES

Yue Wei*, University of Pittsburgh
Ying Ding, University of Pittsburgh

The uptake of targeted therapies has significantly changed the field of medicine. Instead of the "one-fits-all" approach, one aspect is to develop treatments that target a subgroup of patients. In this process, many markers need to be screened, and within each marker, it is necessary to infer treatment efficacy in subgroups and their combinations. For example, for a SNP that separates patients into three subgroups (AA, Aa and aa), one needs to decide whether to

target a single subgroup (e.g., aa) or a combination of subgroups (e.g., {Aa, aa}). In this research, we develop a simultaneous inference procedure to identify subgroups with enhanced treatment efficacy in clinical trials with time-to-event outcomes. Specifically, after establishing a suitable efficacy measure, we provide simultaneous confidence intervals, which appropriately adjust both within- and between-marker multiplicities, for comparing subgroups or combinations of subgroups. Realistic simulations are conducted using true genotype data and various efficacy scenarios to evaluate method performance. This approach contributes in identifying patient subgroups in targeted therapy development.

email: yuw95@pitt.edu

4d. A BASKET TRIAL DESIGN USING BAYESIAN MODEL AVERAGING

Akihiro Hirakawa*, The University of Tokyo
Ryo Sadachi, The University of Tokyo

A single-arm basket trial in oncology often evaluates the response rate of a targeted therapy corresponding to a common molecular characterization in many different cancers. The accurate estimation of the response rate is challenging because sample size of each cancer type is limited and further true response rate is possibly varied depending on cancer types. In this study, we propose a novel Bayesian model averaging (BMA) approach for estimating response rate that accounts for heterogeneity of response rate among cancer types with limited sample size. In the proposed approach, for the cancer types selected based on the results of assessment of response rate homogeneity during the trial, we estimate a posterior promising probability of efficacy and subsequently determine whether the targeted therapy is promising for further testing. We also propose the BMA-based posterior estimator of the response rate. Simulation studies demonstrated that the proposed approach provided better control of the FWER and reduced bias of the response rate than the existing approaches.

email: hirakawa-akihiro@umin.net

4e. OUTCOME WEIGHTED Ψ -LEARNING FOR INDIVIDUALIZED TREATMENT RULES

Mingyang Liu*, University of Minnesota
Xiaotong Shen, University of Minnesota
Wei Pan, University of Minnesota

An individualized treatment rule is often employed to maximize a certain patient-specific clinical outcome based on his/her clinical or genomic characteristics as well as heterogeneous response to treatments. Existing methods such as the partial least squares suffers from the difficulty of indirect maximization of a patient's clinical outcome; while the outcome weighted learning (Zhao et al., 2012) is not robust against perturbation of the outcome. In this article, we propose a weighted ψlearning method to optimize an individualized treatment rule, which is robust against any data perturbation near the decision boundary by seeking the maximum separation. Then we employ a difference convex algorithm to relax the non-convex minimization iteratively based on a decomposition of the cost function into a difference of two convex functions. On this ground, we also introduce a variable selection method for further removing redundant variables for a higher performance. Finally, we illustrate the proposed method by simulations and a lung health study and demonstrate that it yields higher performances in terms of accuracy of prediction of individualized treatment.

email: liux3941@umn.edu

4f. DOMAIN ADAPTATION MACHINE LEARNING FOR OPTIMIZING TREATMENT STRATEGIES IN RANDOMIZED TRIALS BY LEVERAGING ELECTRONIC HEALTH RECORDS

Peng Wu*, Columbia University
Yuanjia Wang, Columbia University

Recently, the estimation of individualized treatment rules (ITRs) in both randomized controlled trials and observational studies (e.g. EHRs) has increasingly received attention. RCTs are often conducted under stringent criterion, limiting the generalizability of ITRs learned from RCT to the broader patient population. Since EHRs document treatment prescription in the real world, transferring information learned from EHRs to RCTs, if done successfully, could bring extra benefit on the performance of ITRs learned from randomized trials alone. We propose a domain adaptation machine learning method to enhance RCT ITRs by leveraging EHRs and achieve domain adaptation from EHRs to RCT through supervised learning feature extraction and projection. In this framework, two machine learning methods including Q-learning and Kernel-Weighted Matched Learning are applied to both observational EHR and randomized trials to accurately estimate ITRs. Simulation studies demonstrate superiority of proposed method in certain scenarios. Lastly, we apply our framework to transfer information learned from EHR of type 2 diabetes patients to improve optimizing individualized insulin therapy in a RCT.

email: pw2394@cumc.columbia.edu

5. POSTERS: CANCER APPLICATIONS

5a. BARCODING OF HEMATOPOIETIC STEM CELLS: APPLICATION OF THE SPECIES PROBLEM

Siyi Chen*, Rice University
Marek Kimmel, Rice University
Katherine Yudeh King, Baylor College of Medicine

We consider estimation of heterogeneity of hematopoietic stem cells (HSCs) which can give rise to all different types of mature blood cells. Cells can be uniquely labeled by insertion of so-called DNA barcodes which is accomplished by transduction of a library of unique DNA sequences. This also allows tracing descendants of the originally labelled cells. The estimation of the size of the target population is very important in assessing the robustness of observations made in experiments. Due to proliferation of the originally barcoded cells, the barcoding experiments will result in multiple replicates of barcodes, and this coincides with the settings of the classical Species Problem. We adopt a uniform prior for the number of 'species' based on the assumption that all barcodes are evenly distributed in the blood system. Then we compare the maximum a posteriori (MAP) estimates i.e. the modes of posterior distributions based on different assumed HSC counts. We also carry out simulations

and plate number computations to determine the number of plates sufficient to achieve statistical power of pooled barcoding experiments.

email: sc52@rice.edu

5b. INFERRING CLONAL EVOLUTION OF TUMORS FROM RNA-seq DATA

Tingting Zhai*, University of Kentucky
Jinpeng Liu, University of Kentucky
Arnold J. Stromberg, University of Kentucky
Chi Wang, University of Kentucky

Many types of tumors are highly heterogeneous containing multiply distinct populations of tumor cells, each with their own complement of somatic mutations. The therapy that targets mutations closer to the trunk may have a better chance of eliminating cancer, while a therapy that targets branch mutations may be less effective. However, to date there are limited methods for inferring clonal evolution of tumor from RNA-seq data. In this poster, we present a new statistical method to reconstruct the evolutionary history and population frequency of the subclonal lineages of tumor cells from RNA-seq measurements. Our method uses a Bayesian nonparametric prior and nested stick-breaking process to allow for evolutional trees of infinite nodes, and to identify cell population frequencies which have the highest likelihood of generating the observed RNA-seq data. Markov chain Monte Carlo method based on slice sampling is incorporated to perform Bayesian inference. Simulations demonstrate that the proposed method reliably recover the phylogenetic chain and population frequency of the subclonal lineages of tumor cells.

email: tzh232@g.uky.edu

5c. MODELING THE EFFECT OF TREATMENT TIMING ON SURVIVAL WITH APPLICATION TO CANCER SCREENING

Wenjia Wang*, University of Michigan
Alexander Tsodikov, University of Michigan

Cancer screening in the population brings life-saving benefit via early diagnosis and treatment, yet it may also cause overdiagnosis and overtreatment. To ensure that benefits outweigh the harms, it is critical to assess the effect of treatment timing implied by the screening and treatment policy on cancer-specific mortality and survival. We formulate a semiparametric mechanistic joint model to study the problem and develop a test for the causal effect of screening. Profile likelihood and the EM algorithm are used for statistical inference. Large-sample properties of proposed estimators are established. The methodology is illustrated by simulation studies and analysis of real data.

email: icywang@umich.edu

5d. MOLECULAR SIGNATURE PREDICTIVE OF SURVIVAL IN METASTATIC CUTANEOUS MELANOMA

Yuna Kim*, Drexel University
Issa Zakeri, Drexel University
Sina Nassiri, The Swiss Institute of Bioinformatics (SIB)

Cutaneous melanomas are classified into subtypes based on their clinicopathological and/or main genomic alterations. Although beneficial for the therapeutic management of the disease, the existing subtypes demonstrate poor correlation with patient survival. By making use of both gene expression profile and clinical data from The Cancer Genome Atlas (TCGA), this study aimed to identify molecular subtypes that are biologically meaningful and clinically relevant in predicting patient survival. Semi-supervised Principal Component (SPC) method was applied to develop and validate a gene expression signature comprised of 1051 genes significantly associated with overall survival in patients with metastatic melanoma. We further validated our gene signature in a subset of the TCGA data that was never used in deriving the signature and showed that its association with overall survival is independent of tumor stage and patient's age. Interestingly, we found that, on average, higher expression of the identified signature is associated with prolonged survival times. Overall, this study provides proof of concept in classifying metastatic melanoma into prognostically distinct subtypes.

email: yk424@drexel.edu

5e. CAN A TUMOR'S TRANSCRIPTOME PREDICT RESPONSE TO IMMUNOTHERAPY?

Shanika A. De Silva*, Drexel University
Sina Nassiri, The Swiss Institute of Bioinformatics (SIB)
Issa Zakeri, Drexel University

Anti-PD-1 immunotherapy has provided numerous clinical benefits to melanoma patients. However, its benefits are often hindered by a 70% resistance rate through mechanisms that remain largely unclear. Previous studies have successfully used the wealth of information in the transcriptomic features of tumors to predict responsiveness to therapy. This study aims at devising a gene signature based on pre-treatment transcriptional profiles to predict response to anti-PD-1 therapy in metastatic melanoma. After the pre-processing of data, a mean-median ratio criterion filter was applied to reduce the number of features. Next, subsets of informative genes were extracted using three differential expression analysis methods. After testing several cross-validated statistical learning algorithms for the purpose of classification with feature selection, it was found that support vector machine outperformed the other classifiers. Consequently, a 138-genes signature, which predicted response with a 75% accuracy, was derived. This signature could assist clinicians in choosing the optimal course of treatment for their patients and aid in the design and execution of future clinical trials.

email: sad345@drexel.edu

5f. SIMILARITY AND DIFFERENCE BETWEEN TELOMERASE ACTIVATION AND ALT BASED ON THE THEORY OF G-NETWORKS AND STOCHASTIC AUTOMATA NETWORKS

Katie Kyunghyun Lee*, Rice University
Marek Kimmel, Rice University

The oldest, but still the most fundamental, question is whether the mathematical approaches can precisely model and interpret gene regulatory networks with genetic interactions that control gene expression. In this research, we propose that two stochastic queueing models, G-Networks and Stochastic Automata Networks are

useful to identify genes important in the disease of interest. Further, our purpose is to infer their correlation by obtaining both stationary and transient distributions of the system. Here, we apply these methods to a network including genes particularly related to telomeres. In cancer cells, their lengths are stabilized, thereby allowing continual cell replication by two mechanisms: activation of telomerase and Alternative Lengthening of Telomeres (ALT). Our analysis detects five statistically significant genes in cells with either mechanism. Moreover, we introduce a novel algorithm to show how the correlation between two genes of interest varies not only according to each mechanism but also to each cell condition. This study expands our existing knowledge of genes associated with telomere maintenance and the similarities/differences between telomerase and ALT.

email: kl25@rice.edu

5g. OPTIMIZATION OF MOMENT-BASED INTENSITY-MODULATED RADIATION THERAPY (IMRT) TREATMENT PLAN

Wanxin Chen*, Temple University
Abraham Abebe, Temple University

Intensity-Modulated Radiation Therapy (IMRT) is an advanced technique used in Cancer Radiotherapy. When a required amount of radiation is given to cancerous tumors of patients, radiation fields are expanded, potentially resulting in collateral damages to surrounding healthy tissues. The dose-volume histogram (DVH) is used to evaluate the effectiveness of a treatment plan. Literature concluded the approach of employing two and three moments formulation of random variable to approximate desired DVHs, suggesting improvements could be made. The goal of our work was to improve predictions about dosage requirements for future radiation sessions based on the original real-world DVH curves. We provided gradient information in Matlab's constrained minimization function, added an additional phase onto the algorithm with two and three moments and further shrunk the moments' bounds after getting the Pareto-optimal plan from the two-phase approach. This has significantly improved the accuracy and predictability on the DVH, which in turns, minimized treatment risk. We also determined that three moments were optimal by comparing reference and computed DVH curves.

email: idjoannachen@gmail.com

6. POSTERS: CLINICAL TRIALS

6a. TWO-STAGE ADAPTIVE ENRICHMENT DESIGN FOR TESTING AN ACTIVE FACTOR

Rachel S. Zahigian*, University of Florida
Aidong Adam Ding, Northeastern University
Samuel S. Wu, University of Florida
Natalie E. Dean, University of Florida

When the effect of an intervention varies across subpopulations, testing for a main effect in the combined population may be inefficient, and testing each subpopulation separately may require a large sample size. Investigators may consider a narrowed objective and test whether the intervention is an active factor. We define a factor as active if its effect is non-zero in at least one subpopulation. This composite hypothesis requires a smaller overall sample size than testing each subpopulation separately and is more robust than main effect testing when interactions are present. To implement active factor testing, we propose a two-stage adaptive enrichment strategy using a play-the-winner design. Using numerical simulations, we determine that the most powerful analytical strategy is to use a noncentral chi-square test statistic in the second stage and combine the p-values from the two stages using a weighted Fishers combination. Simulations indicate that our proposed test generally improves upon a single-stage trial of equivalent design. Active factor testing may be useful for exploratory and pilot testing and can be used to screen for gene-treatment interactions.

email: rzahigian@ufl.edu

6b. COVARIATE ADJUSTMENT FOR CONSIDERING BETWEEN-TRIAL HETEROGENEITY IN CLINICAL TRIALS USING HISTORICAL DATA FOR EVALUATING THE TREATMENT EFFICACY

Tomohiro Ohigashi*, Tokyo University of Science
Takashi Sozu, Tokyo University of Science
Jun Tsuchida, Tokyo University of Science

It is difficult to enroll an adequate number of participants in clinical trials involving children or patients with rare diseases. In order to deal with this issue, a Bayesian approach of incorporating historical control data from previous studies into the current study was proposed for evaluating a new treatment's efficacy. However, the type I error rate tends to be inflated if heterogeneity exists between the historical and new data. We evaluated the operating characteristics of the Bayesian hierarchical model that incorporates patient-level covariates. When the covariates that yielded a difference in patient characteristics between the historical and current trials were correctly specified, the type I error rates were adequately controlled to below the significance level, and the powers were almost 5% higher than those of the separate method (Fisher's exact test without using the historical data). We also show the results obtained when the variables that were strongly correlated with the true covariates yielding the between-trial heterogeneity were specified.

email: 4417606@ed.tus.ac.jp

6c. A COMPARISON OF STATISTICAL METHODS FOR TREATMENT EFFECT TESTING AND ESTIMATION WITH POTENTIAL NON-PROPORTIONAL HAZARDS

Jing Li*, Indiana University
Qing Li, Merck & Co., Inc.
Amarjot Kaur, Merck & Co., Inc.

In current practice of clinical trials, log-rank test and Cox Proportional Hazards (Cox PH) models are used as gold standards for testing and estimating treatment effect. A crucial assumption for the validity of Cox PH and efficiency of log-rank test is proportional hazards. However, non-proportional hazards (NPH) are not uncommon especially in immuno-oncology trials. There is no general consensus amongst industry and regulatory agencies on the choice of optimal methods with possible NPH. Although much research has been conducted in literature on both testing and estimation of treatment effect, each alternative method has its own advantages and limitations. In this project, we considered several alternative approaches, including the emerging max-combo, conventional and weighted

log-rank, weighted Kaplan-Meier tests, restricted mean survival time method, and Cox PH model with single change point. We compared their performance in testing and/or estimating treatment effect under a wide range of scenarios for proportional and non-proportional hazards with the goal of identifying robust method(s) across different simulation situations examined.

email: JL204@iu.edu

6d. DOSE-FINDING METHOD FOR MOLECULARLY TARGETED AGENTS INCORPORATING THE RELATIVE DOSE INTENSITY IN PHASE I ONCOLOGY CLINICAL TRIALS

Yuichi Tanaka*, The University of Tokyo
Takashi Sozu, The University of Tokyo
Akihiro Hirakawa, The University of Tokyo

The purpose of phase I oncology clinical trials is to determine the recommended phase 2 dose (RP2D) for further testing. For a cytotoxic agent, RP2D corresponds to the maximum tolerated dose (MTD), which is defined as the highest dose administered to patients with clinically acceptable toxicity. MTD is often determined based solely on the toxicity data obtained at the first treatment cycle. However, MTD-based determination of RP2D for molecularly targeted agents (MTAs) may not be adequate because MTAs have late-onset and/or cumulative toxicity. To evaluate the longitudinal toxicities of MTAs, we proposed a new method to determine RP2D, which takes into account the relative dose intensity (RDI). RDI is generally defined as the ratio of effectively administered to theoretically administered cumulative dose. Further, we also proposed a dose-finding method to identify novel RP2D for MTAs, based on the modified 3+3 design, where the RDI was evaluated at each cycle. The new RP2D tended to be lower than the MTD. The number of participants required for the study were found to be relatively lower than those required by the traditional method.

email: 4417618@ed.tus.ac.jp

6e. USE OF QUADRATIC INFERENCE FUNCTION FOR ESTIMATION OF MARGINAL INTERVENTION EFFECTS IN CLUSTER RANDOMIZED TRIALS

Hengshi Yu*, University of Michigan
Fan Li, Duke University
Elizabeth Louise Turner, Duke University

Cluster randomized controlled trials (c-RCTs) are trials that randomize clusters, but measure outcomes on individuals. Marginal intervention effects are commonly of interest for their population-averaged interpretation. Such effects are typically estimated using the generalized estimating equation (GEE) approach for the correlated nature of outcomes measured on individuals from the same cluster. An alternative approach is the quadratic inference function (QIF) approach. Evidence from the literature suggests that QIF can provide an efficiency improvement over GEE in the longitudinal data setting. There is limited evidence of their potential benefits for the correlated data in c-RCTs. In this paper, we apply QIF and GEE in the estimation of the marginal intervention effects in c-RCT with continuous outcomes at one follow-up time point. We concentrate on c-RCTs with a large number of clusters and approximately the same cluster sizes. We compare QIF and GEE through simulation studies and provide three novel theorems about equivalence of point estimation from QIF and GEE. We demonstrate the two methods using data from a real c-RCT in Kenya.

email: hengshi@umich.edu

6f. ONGOING CLINICAL TRIAL FORECAST AND MONITORING TOOL: A STATISTICAL FRAMEWORK

Shijia Bian*, Biogen
Jignesh Parikh, Sema4
Tanya Cashorali, Biogen and TCB Analytics
Mike Fitzpatrick, Biogen
Murray Abramson, Biogen
Feng Gao, Biogen

A feasible approach to assist decision making for cross-functional teams which involves in Clinical Trial (CT)-associated activities is widely missing in healthcare. The integration of statistical method, machine learning and data management is one potential approach to meet this unmet need. The merge of these fields, which can reduce boundary through information fusion and aid in data mining for making data-driven decisions, is an innovative area has been overlooked by many well-established Biostatistics sectors. In this work, we proposed a framework to support ongoing CT enrollment forecast and monitoring, with a purpose to enhance the confidence in making timely consistent and data-driven decisions. We started with kitchen-sink approach for model selection. Benefiting from Bayesian methodology, the final statistical model was a robust predictive model composed with timely-updated estimators inferred from the real-time CT data. This framework can perform scenario analysis to mimic different real-word scenarios. Tableau and Shiny were utilized for visualization. Broader usage of this tool could be explored after being tuned with more CTs.

email: shijia.bian@biogen.com

6g. PARAMETER ESTIMATION USING INFLUENTIAL EXPONENTIAL TILTING IN CASE OF DATA MISSING AT RANDOM AND NON-IGNORABLE MISSING DATA

Kavita A. Gohil*, Georgia Southern University
Hani M. Samawi, Georgia Southern University
Haresh Rochani, Georgia Southern University
Lili Yu, Georgia Southern University

Inference with incomplete data is one of the challenging tasks in clinical trials when missing data can be related to the condition under study. The benefits of randomization are compromised by incomplete and missing data. Multiple imputation is a valid method of treating missing data under the assumption of MAR (Missing at random) with the use of sensitivity analysis to adjust for departure from the MAR missingness. Parameter estimation with non-ignorable missing (missing not at random) data is even more challenging to handle and extract useful information. The missing values if accompanied by measurement error would make the situation worse. The goal of this research is to modify the multiple imputation methods for dealing with missing not at random data with measurement error using the influential exponential tilting approach. To assess the efficiency of the proposed method on

misclassification error, a fully Bayesian method is to be used. Intensive simulation study will be conducted to get an insight into the proposed approach and compare it with other existing approaches. Theoretical justification is provided as well.

email: kg03393@georgiasouthern.edu

7. POSTERS: DIAGNOSTICS/AGREEMENT

7a. A NON-INFERIORITY TEST FOR COMPARING TWO PREDICTIVE VALUES OF DIAGNOSTIC TESTS

Kanae Takahashi*, Osaka City University
Kouji Yamamoto, Yokohama City University

Diagnostic tests are important for the early detection and treatment of disease in medicine. The positive predictive value (PPV) and the negative predictive value (NPV) describe how good the test is at predicting abnormality, and these are used for quantifying the diagnostic ability of the test. The PPV is the probability of disease when the diagnostic test result is positive, and the NPV is the probability of no disease when the diagnostic test result is negative. There are several methods to test the equality of predictive values in paired designs. Some researchers utilize the multivariate central limit theorem and the delta method, and others use regression frameworks. These methods were developed for superiority trial. Although the testing of non-inferiority may be adequate in some cases due to the simplicity of the examination and the small invasiveness, the non-inferiority test for predictive values has not been proposed yet. In this study, we propose a non-inferiority test for conducting a clinical trial that investigates the non-inferiority of predictive values in paired designs.

email: takahashi.kanae@med.osaka-cu.ac.jp

7b. IMPROVING INFERENCE ON DISCRETE DIAGNOSTIC TESTS WITHOUT A GOLD STANDARD

Xianling Wang*, University of Pittsburgh
Gong Tang, University of Pittsburgh

Discrete diagnostic tests such as tumor grade in cancer are usually important prognostic factors but often suffer from intra-rater and inter-rater reproducibility. With at least three independent readings, the prevalence and classification rates of independent raters are estimable up to a permutation of labels for the unknown truth under a latent class model (Kruskal, 1977; Dawid, 1979). Violation of the conditional independence assumption on the independent raters leads to biased estimates (Vacek, 1985). Although the raters may rate the study subjects independently, dependence can be induced by certain nature of the common scoring system used. Here a new nesting latent class method is proposed to deal with such induced dependence and the likelihood ratio test is used to check the violation of conditional independence. With available auxiliary variable, the model is further extended to provide global identification of model parameters. The methods are illustrated by an analysis of tumor grade reading data from a joint study of the National Surgical Adjuvant Breast and Bowel Project (NSABP) and the Genomic Health Inc.

email: xiw118@pitt.edu

7c. APPLICATIONS OF GENERALIZED KULLBACK-LEIBLER DIVERGENCE AS A MEASURE OF MEDICAL DIAGNOSTIC AND CUT-POINT CRITERION FOR K-STAGES DISEASES

Chen Mo*, Georgia Southern University
Hani M. Samawi, Georgia Southern University
Jingjing Yin, Georgia Southern University
Haresh D. Rochani, Georgia Southern University
Xinyan Zhang, Georgia Southern University
Jing Kersey, Georgia Southern University

Recently, Kullback-Leibler divergence measure (KL), which captures the disparity between two distributions, has been considered as an index for determining the diagnostic performance of biomarkers. Our study investigates variety of applications of the generalized KL divergence (GKL) in medical diagnostics, including overall measures of rule-in and rule-out potential and proposes an optimization criterion based on KL divergence for cut-point selection for k-stage disease. Moreover, the study links the KL divergence with some common Receiver Operating Characteristic (ROC) measures and presents analytically and numerically the relations in situations of one cut-point as well as multiple cut-points. Furthermore, the graphical application and interpretation of GKL divergence, which is referred as the information graph, is discussed. A comprehensive data analysis of the real data example to illustrate the proposed applications is provided.

email: cm06957@georgiasouthern.edu

7d. A NONPARAMETRIC PROCEDURE FOR COMPARING DEPENDENT KAPPA STATISTICS

Hanna Lindner*, University of Pennsylvania
Phyllis Gimotty, University of Pennsylvania
Warren Bilker, University of Pennsylvania

Quantifying agreement using the kappa statistic is often of interest as it serves as a measure of inter-rater reliability when the judgement process is a binary rating. We consider a scenario with a set of observations from multiple groups (between factor) that are given a consensus rating and at least one other independent rating (within factor). Interest centers on how the kappa values, computed using the consensus rating, differ across levels of the within factor, the between factor, and if a between by within interaction is present. A key feature of the data is that the kappa statistics are dependent, since they are all based on the consensus ratings. This method allows us to nonparametrically test for a between, within, and interaction effect using permutation tests and bootstrap, while adjusting for the correlation structure. This method, KAPPANOVA, is analogous to a two-way ANOVA, but considers differences in kappa, rather than means. It is a modification of the CORANOVA procedure (Bilker, 2004), which compares dependent correlations. The method is applied to biomarker data, where biomarkers act as raters and agreement on response to treatment is the consensus rater.

email: hlindner@pennmedicine.upenn.edu

7e. EVALUATING DIFFERENT APPROACHES TO CLASSIFY PATIENTS OF VECTOR TRANSMITTED VIRAL INFECTIONS USING SYMPTOM INFORMATION

Ana Maria Ortega-Villa*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Sally Hunsberger, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Wenjuan Gu, Frederick National Laboratory for Cancer Research sponsored by the National Cancer Institute, National Institutes of Health
Keith Lumbard, Frederick National Laboratory for Cancer Research sponsored by the National Cancer Institute, National Institutes of Health
Jesús Sepúlveda-Delgado, Hospital Regional de Alta Especialidad Ciudad Salud. Tapachula, Chiapas
Pablo F. Belaunzaran-Zamudio, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico

Detection and identification of vector transmitted viral infections is of critical importance for health care providers in the event of an outbreak. However, the co-circulation of viruses like Zika and Dengue presents challenges in distinguishing between them. In this study, we evaluated the classification accuracy of various statistical methodologies including logistic and multinomial regression, discriminant correspondence analysis and decision trees based on symptom data alone. In addition, we propose a group of symptoms that may aid in discriminating PCR confirmed Zika and Dengue patients from other unknown viral infections. This analysis was performed using data from the La Red Mexican Emerging Infectious Disease Clinical Research Network study, a prospective observational longitudinal multicentric study in the Tapachula area in Mexico.

email: ana.ortega-villa@nih.gov

8. POSTERS: ADAPTIVE DESIGN/EXPERIMENTAL DESIGN

8a. A UTILITY-BASED SEAMLESS PHASE I/II TRIAL DESIGN TO IDENTIFY THE OPTIMAL BIOLOGICAL DOSE FOR TARGETED AND IMMUNE THERAPIES

Yanhong Zhou*, University of Texas MD Anderson Cancer Center
J. Jack Lee, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center

In the era of targeted therapy and immunotherapy, the objective of dose finding is to identify the optimal biological dose (OBD). We develop a utility-based seamless (UBS) phase I/II design to find the OBD for immunotherapy. We show that the utility approach is flexible and more general in the sense that it contains the existing toxicity-efficacy trade-off methods as special cases. To accommodate the common case that efficacy of immunotherapy may take a relatively long time to be ascertained, we leverage the short-term endpoint, e.g., immune activity or other biological activity of targeted agents, and use it to predict the unobserved efficacy outcome to facilitate the real-time decision making. During the trial, in light of accumulating efficacy and toxicity, we continuously update the posterior estimate of the utility for each dose after each cohort and use it to direct the dose assignment and selection. Simulation study shows that the UBS identifies the OBD accurately and significantly shortens the trial duration. A user-friendly software to implement the UBS design is also available.

email: yzhou13@mdanderson.org

8b. AN ADAPTIVE BIOMARKER-DRIVEN PHASE II DESIGN

Jun Yin*, Mayo Clinic
Daniel Kang, University of Iowa
Qian Shi, Mayo Clinic

The paradigm for cancer clinical trials has shifted towards individualized treatment due to the blooming discoveries of biomarkers and targeted agents. I will introduce a novel statistical framework for designing and conducting such umbrella/basket trials. The proposed design allows the parallel cohorts to be dynamic, so that at interim analysis investigators can 1) eliminate or graduate marker/treatment cohorts, and 2) cluster marker/treatment cohorts to improve the efficiency and power to identify promising targeted agents. The design is calibrated with respect to specific error rates. Extensive simulations are conducted to assess the performance of the proposed design with comparison to sequentially conducted single-arm trials, and illustrate it with an NCTN coordinated trial in brain metastases.

email: vivien.jyin@gmail.com

8c. A SIGNATURE ENRICHMENT DESIGN WITH BAYESIAN ADAPTIVE RANDOMIZATION FOR CANCER CLINICAL TRIALS

Fang Xia*, University of Texas MD Anderson Cancer Center
Stephen L. George, Duke University School of Medicine
Jing Ning, University of Texas MD Anderson Cancer Center
Liang Li, University of Texas MD Anderson Cancer Center
Xuelin Huang, University of Texas MD Anderson Cancer Center

Clinical trials in the era of precision cancer medicine aim to identify and validate biomarker signatures which can guide the assignment of individually optimal treatments to patients. Here we propose a cross-validated signature enrichment design combined with Bayesian response-adaptive randomization to address these objectives. We evaluate the performance of the design based on four criteria utilizing the benefits and losses for individuals both inside and outside of the clinical trial. The proposed design allows more patients to receive optimal personalized treatments, thus yielding a higher overall response rate. This design can identify therapies that are globally beneficial as well as treatments that are effective only in a sensitive subset. Simulation studies demonstrate the advantages of the proposed design over alternative designs. The approach is illustrated by an example based on an actual clinical trial in non-small-cell lung cancer.

email: fang.katrina.xia@gmail.com

8d. BLINDED SAMPLE SIZE RE-ESTIMATION IN COMPARATIVE CLINICAL TRIALS WITH OVER-DISPERSED COUNT DATA: INCORPORATION OF MISSPECIFICATION OF THE VARIANCE FUNCTION

Masataka Igeta*, Hyogo College of Medicine
Shigeyuki Matsui, Nagoya University School of Medicine

In clinical trials with over-dispersed count data, the sample size calculated for comparing event rates across treatment arms may heavily depend on nuisance parameters, including the overall event rate and the working and true variance functions in negative binominal or quasi-Poisson models. In this paper, we propose

a method for blinded sample size re-estimation anticipating misspecifications of the variance functions specified at the planning stage. For treatment comparison, we utilize a Wald-type test with a sandwich-type robust variance estimator under a quasi–Poisson model. The sample size is updated based on estimates of the nuisance parameters obtained by analyzing interim blinded data. Our method includes several existing methods, which require correct specification of the working variance function, as special cases. Simulation studies demonstrated our method can control the Type I error rate and power better than conventional ones based on model-based standard errors when the sample size is relatively large. An application to a clinical trial to evaluate the treatment effect on prevention of exacerbations of chronic obstructive pulmonary disease is provided.

email: ma-igeta@hyo-med.ac.jp

8e. AN EXPLORATION OF OPTIMAL DESIGN ROBUSTNESS IN NONLINEAR MODELS

Ryan Jarrett*, Vanderbilt University
Matthew S. Shotwell, Vanderbilt University

Optimal designs identify a set optimal factor levels that maximize the information gain for a given sample. Information gain is typically measured by some function of the Fisher Information Matrix (FIM), which depends on the unknown parameter values for nonlinear models. While the optimality criterion can be evaluated at best guess values in a locally optimal design, researchers may try to increase design robustness by averaging the design information over a prior distribution of parameter values (ED-optimal), or finding the optimal design at the worst combination of parameter values (minimax-optimal). We investigate how robust these approaches are to modeling errors that are likely to occur in pharmacokinetic studies. Specifically, we consider design robustness to misspecification of the model parameters, parameter prior distribution, and response function. We also examine how well each approach manages expectations regarding information gain by considering the variability and skew of the optimality criterion distribution. Finally, we consider the robustness benefits of supplementing optimal designs with measurements at sub-optimal factor levels.

email: ryan.t.jarrett@vanderbilt.edu

9. POSTERS: BAYESIAN METHODS

9a. CONSTRUCTING A PRIOR FOR THE CORRELATION COEFFICIENT USING EXPERT ELICITATION

Divya R. Lakshminarayanan*, Baylor University
John W. Seaman Jr., Baylor University

Prior elicitation is defined as formulating an expert's beliefs about one or more uncertain quantities into a joint probability distribution, and is often used in Bayesian statistics for specifying prior distributions for parameters in the data model. However, there is limited research on eliciting information about dependent random variables. This research concerns constructing a prior distribution for the correlation between dependent parameters using expert elicitation. We will consider methods for modeling the correlation between two Bernoulli success probabilities, and methods for modeling the correlation for a bivariate normal distribution.

email: drlakshm93@gmail.com

9b. CONSISTENT BAYESIAN JOINT VARIABLE AND DAG SELECTION IN HIGH DIMENSIONS

Xuan Cao*, University of Cincinnati
Kshitij Khare, University of Florida
Malay Ghosh, University of Florida

Motivated by the eQTL analysis, we consider joint sparse estimation of the regression coefficient matrix and the error covariance matrix in a high-dimensional multivariate regression model. The error covariance matrix is modeled via Gaussian directed acyclic graph (DAG) and sparsity is introduced in the Cholesky factor of the inverse covariance matrix, while the sparsity pattern in turn corresponds to specific conditional independence assumptions on the underlying variables. By considering a flexible and general class of these `DAG-Wishart' priors with multiple shape parameters on the space of Cholesky factors and a spike and slab prior on the regression coefficients, we establish the joint selection consistency for both the variable and the underlying DAG when both the number of predictors and the dimension of the covariance matrix are allowed to grow much larger than the sample size. We demonstrate our theoretical results through a marginalization-based collapsed Gibbs sampler that offers a computationally feasible and efficient solution for exploring the sample space.

email: xuan.cao@uc.edu

9c. BAYESIAN APPROACH FOR JOINT MODELLING OF LONGITUDINAL AND TIME TO EVENT DATA

Zeynep Atli*, Mimar Sinan Fine Arts University
Mithat Gönen, Memorial Sloan Kettering Cancer Center
Gülay Basarir, Mimar Sinan Fine Arts University

Joint models of longitudinal and time to event data provide simultaneous parameter estimation for longitudinal and time to event process through shared random effects. The estimation of parameter is usually based on maximizing the joint likelihood function. Maximization of joint likelihood function has to deal with high-dimensional integrals arising from random effects. To better handle the computational, we adopt a Bayesian approach using the Metropolis-Hasting algorithm combined with Gibbs sampling for parameter estimation. We will present the details of our Bayesian model and compare its operating characteristics with the likelihood approach.

email: zeynep.atli@msgsu.edu.tr

9d. RATIONAL DETERMINATION OF THE BORROWING RATE IN BAYESIAN POWER PRIOR MODELS

Mario Nagase*, AstraZeneca
Shinya Ueda, AstraZenecaKK
Mitsuo Higashimori, AstraZenecaKK
Katsuomi Ichikawa, AstraZenecaKK
Jim Dunyak, AstraZeneca
Nidal Al-Huniti, AstraZeneca

Historical data is accumulating rapidly, offering the promise of better-informed future clinical trials through Bayesian methods. On the other hand, past trials are not specifically designed for this purpose, so it is difficult to fully incorporate them into current analyses. Various methods for partially borrowing historical data, e.g., the Power Prior Model (PPM), have been developed, but determination of borrowing rate is still a complex question. One method is to determine the rate a priori based on the background similarity, but assessing this similarity is challenging. Many of the other methods based on frequentist statistics lead to very low borrowing rates and provide limited value. The PPM has a hierarchical Bayes interpretation, leading to a relationship between variance and borrowing rate. We propose an approach that directly evaluate the variance by integrating out unnecessary latent parameters. It shows a realistic estimate for the borrowing rate comparable to a subjective similarity approach. As an example, applying the new method to design of a Japan/global bridging study results in a borrowing rate of 24.6%, while a subjective approach leads to around 30%.

email: mario.nagase@astrazeneca.com

9e. A SEMIPARAMETRIC APPROACH FOR ESTIMATING A BACTERIUM'S WILD-TYPE DISTRIBUTION: ACCOUNTING FOR CONTAMINATION AND MEASUREMENT ERROR (BayesACME)

Will A. Eagan*, Purdue University
Bruce A. Craig, Purdue University

Antimicrobial resistance is a major challenge to modern medicine and of grave concern for public health. To monitor resistance, agencies analyze "drug/bug" collections of clinical assay results. It is common to assume the underlying distribution of results is a mixture of two subpopulations: a wild-type distribution on the left consisting of bacterial strains susceptible to the drug and resistant mutants on the right. The goal is to estimate the wild-type distribution and its prevalence. The most common assay involves placing a bacterium and two-fold concentrations of a drug in a well of a 96-well plate and seeing if there is growth. This choice of two-fold concentrations enforces interval-censored results. Various estimation methods have been proposed that account for this censoring and contamination by the mutant distribution. None of the methods, however, account for the inherent variability of the assay, which can have a profound effect on the estimated distribution. We propose a Bayesian semiparametric method to address this limitation. The improved estimation of the wild-type distribution is demonstrated through a series of simulation studies.

email: weagan@purdue.edu

9f. A BAYESIAN MIXTURE MODEL TO ESTIMATE THE EFFECT OF AN ORDINAL PREDICTOR

Emily Roberts*, University of Michigan
Lili Zhao, University of Michigan

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

email: ekrobe@umich.edu

10. POSTERS: BAYESIAN OMICS/LATENT BAYES MODELS

10a. INFERRING GENE NETWORKS WITH GLOBAL-LOCAL SHRINKAGE RULES

Viral V. Panchal*, Augusta University
Daniel F. Linder, Augusta University

Inferring gene regulatory networks from high-throughput 'omics' data has proven to be a computationally demanding task of critical importance. Indeed, as the number of parameters that require estimation grows quickly, usually much larger than the sample size, most standard methods breakdown and common strategies are typically based on regularization of classical objective functions, like penalized likelihoods. We propose a Bayesian hierarchical model to reconstruct gene regulatory networks from time series gene expression data; i.e., as common in perturbation experiments of biological systems. The proposed methodology utilizes global-local shrinkage priors for posterior selection of regulatory edges and relaxes the common normal likelihood assumption to allow for heavy-tailed data, as was shown in several of the cited references to severely impact network inference. We use an approach based on Hamiltonian Markov chain Monte-Carlo, via implementation in the Stan language, for efficient posterior computation. We demonstrate the performance of our approach in a simulation study and compare it with existing methods on real data from a T-cell activation study.

email: vpanchal@augusta.edu

10b. BAYESIAN KINETIC MODELING FOR TRACER-BASED METABOLOMIC DATA

Xu Zhang*, University of Kentucky
Andrew N. Lane, University of Kentucky
Arnold Stromberg, University of Kentucky
Teresa W-M. Fan, University of Kentucky
Chi Wang, University of Kentucky

Kinetic modeling of the time dependence of metabolite concentrations including the stable isotope labeled species is an important approach to simulate metabolic pathway dynamics. It is also essential for quantitative metabolic flux analysis using tracer data. However, as the metabolic networks are complex including extensive compartmentation and interconnections, the parameter estimation for enzymes that catalyze individual reactions needed for kinetic modeling is challenging. We propose a Bayesian approach that leverages existing expert-constructed kinetic models for specifying an informative prior distribution for kinetic parameters. This prior knowledge prioritizes regions of parameter space that encompass the most likely parameter values, thereby facilitating robust parameter estimation. A component-wise adaptive Metropolis algorithm is used to generate the posterior samples of the kinetic parameters. Simulation studies using defined networks are used to test the performance of this algorithm under conditions of variable noise.

email: xzh323@g.uky.edu

10c. A BAYESIAN APPROACH FOR FLEXIBLE CLUSTERING OF MICROBIOME DATA

Yushu Shi*, University of Texas MD Anderson Cancer Center
Liangliang Zhang, University of Texas MD Anderson Cancer Center
Kim-Anh Do, University of Texas MD Anderson Cancer Center
Robert Jenq, University of Texas MD Anderson Cancer Center
Christine Peterson, University of Texas MD Anderson Cancer Center

With the advent of next generation sequencing technologies, researchers are now able to cheaply and comprehensively analyze microbial communities as never before. Because of the highly diverse nature of the human microbiome, unsupervised clustering is often used to identify subjects with distinct microbial populations. These clusters can then be assessed for associations to sample characteristics of interest. In this work, we recognize that microbiome data sets often contain ``noise'''' OTUs (operational taxonomic units) that hinder successful clustering, and propose a method that can simultaneously select informative OTUs and cluster observations using a Dirichlet process mixture of Dirichlet Multinomial distributions. Unlike previous methods, our proposed method can learn the number of clusters from the data and does not require pre-specifying the number of groups. We test the performance of our method on simulated data and illustrate its application to a real dataset.

email: yshi7@mdanderson.org

10d. LATENT SCALE PREDICTION MODEL FOR NETWORK VALUED COVARIATES

Xin Ma*, Emory University
Suprateek Kundu, Emory University
Jennifer Stevens, Emory University

Network valued data commonly arise in areas such as neuroimaging, genetics and social sciences. Although networks contain rich information, there have been limited advances in regression approaches involving network valued covariate. The high dimensionality of the networks often results in models with inflated number of parameters leading to computational burden and inaccurate estimation. Alternative approaches seek to reduce network dimension and then use the low-rank structure in prediction. This class of methods often lack interpretability and exploratory value. In this work, we develop a novel two stage Bayesian framework to find a node-specific low-rank representation for the network covariates and then use a flexible regression framework for prediction. The approach results in a dramatic reduction in the number of regression parameters and is able to maintain interpretability at the node level. The computation is realized through an efficient EM algorithm. We evaluate our performance in prediction and inference in simulations and posttraumatic stress disorder application seeking to predict a clinical outcome based on structural connections in the human brain.

email: xin.ma@emory.edu

10e. BAYESIAN ESTIMATION IN LATENT VARIABLE ANALYSIS IN Mplus AND WinBUGS

Jinxiang Hu, University of Kansas Medical Center
Lauren Clark*, University of Kansas Medical Center
Byron Gajewski, University of Kansas Medical Center

Background: Patient Reported Outcome (PRO) is gaining more attention in patient-centered health outcomes research and quality of life studies. Latent variable analysis (LVA) is the best way to analyze PRO. Bayesian estimation is flexible in estimating complex models, especially with small sample size. Popular software for LVA with Bayesian estimation includes Mplus and WinBUGS. However, comparison of the results of the two software, especially using DIC as the model fit index, is rarely examined. To fill this gap, we conducted a simulation study. Method: Our preliminary study focused on a one factor model. Simulation conditions include: sample size (50, 100, 200), intercept difference (0, 1), and slope difference (.25, .5, .75). Each condition was iterated 100 times. The homogeneous model, the independent model, and the hierarchical model were fit to each data set generated. The DIC was used as the model fit criterion. Results: Both Mplus and WinBUGS had good parameter recovery. Mplus had high selection rate of the true model than WinBUGS. We also analyzed the clinical data of breast cancer patient assessment of mammography services as the empirical example.

email: jhu2@kumc.edu

10f. A BAYESIAN FACTOR MODEL FOR HEALTHCARE RANKINGS: APPLICATIONS IN ESTIMATING COMPOSITE MEASURES OF QUALITY

Stephen Salerno*, University of Michigan
Lili Zhao, University of Michigan
Yi Li, University of Michigan

Due to increased public interest in the reporting of healthcare metrics, national quality initiatives have focused on greater transparency and interpretability. Ranking statistics are of vital importance to public reporting efforts as they provide an intuitive means of conveying information to patients. Rankings are often based on composite measures, which aggregate information from multiple complex metrics. As these

individual measures arise from different underlying distributions, with different scales, and missing observations, computation becomes particularly challenging. We offer a flexible, rank-based Bayesian factor model for the estimation of such composite observation ranks. We employ the parsimony of Bayesian MCMC for semi-parametric copula estimation and simultaneous missing imputation while resolving identifiability issues with factor rotation. Through this method, the factor scores can be viewed as a scale-free surrogate for estimating the underlying order statistics of our joint distribution. We show in simulation and a real data example that this method can outperform other ad-hoc deterministic or parametric model-based aggregation approaches in certain settings.

email: salerno1212@gmail.com

10g. A BAYESIAN LATENT VARIABLE MODEL FOR INFLAMMATORY MARKERS AND BIRTH OUTCOMES IN SEYCHELLES

Alexis E. Zavez*, University of Rochester
Alison J. Yeates, Ulster University
Edwin van Wijngaarden, University of Rochester
Sally W. Thurston, University of Rochester

The balance of T-helper 1 (Th1) and T-helper 2 (Th2) cytokines may play an important role in maternal immune response throughout pregnancy, thereby potentially affecting birth outcomes. Traditionally, Th1 and Th2 cytokines are either modeled individually, or are summed within Th subsets. However separate models cannot account for joint effects, while summing makes very particular assumptions about combined effects. As an alternative, we propose a flexible Bayesian model with latent Th1 and Th2 variables that accounts for cytokine measurement error. We allow for inclusion of additional inflammatory markers that may be associated with birth outcomes. In an auxiliary MCMC step, we handle the large number of cytokine measurements that are below the lower limit of detection. We apply our model to data from the Seychelles Child Development Study, where twelve inflammatory markers including four Th1 and three Th2 cytokines were measured at 28-week's gestation.

email: alexis_zavez@urmc.rochester.edu

11. POSTERS: GENOMICS/PROTEOMICS

11a. A NONNEGATIVE MATRIX FACTORIZATION METHOD FOR RANK NORMALIZED DATA

Danielle Dematteis*, The College of New Jersey
Michael F. Ochs, The College of New Jersey

We present a method for performing non-negative matrix factorization (NMF) motivated by the problem of normalizing omics data from diverse platforms. Recent work has suggested that rank (quantile) normalization can provide both comparisons to a baseline and robust discovery of differences between observations. Rank approaches should also be maximally robust to use of different platforms, providing a method to compare data gathered with different technologies and at different times. However, such normalization is incompatible with the application of current NMF methods, which is unfortunate given the success of NMF in identifying differences in biological process activities between individuals. Here we present a stochastic NMF method that operates on rank normalized matrices to recover underlying patterns in omics data. Further, we demonstrate the effect of different metrics on the ability of the method to recover meaningful patterns in the data and explore the limitations of the method when applied to gene expression data. In addition, through simulations, we explore the stability of the matrix factorization under increasing noise levels.

email: ochsm@tcnj.edu

11b. SEMIPARAMETRIC METHOD IN RNA-Seq DIFFERENTIAL EXPRESSION ANALYSIS INCORPORATING UNCERTAINTY OF ABUNDANCE ESTIMATES

Anqi Zhu*, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Michael I. Love, University of North Carolina, Chapel Hill

Ideal statistical testing procedures for RNA-Seq would control for technical biases and biological and technical variation, and also incorporate the uncertainty of abundance estimates when testing for differential expression (DE) across conditions. Popular methods for RNA-Seq DE analysis fit a parametric model to the counts of reads for each gene or transcript. If the parametric model does not fit the data well, the method may have inflated false discovery rates (FDR). Previous work showed that nonparametric algorithms for RNA-Seq DE analysis may have better control of FDR. While some parametric models have been proposed to incorporate the uncertainty of abundance estimates into the DE testing, existing nonparametric methods do not incorporate the inferential uncertainty of the observations, which may lead to an inflated FDR. We propose a semiparametric proportional odds ordinal regression applied to inferential replicate datasets, which both accounts for nuisance covariates and the uncertainty of the abundance estimates. Comparing to popular DE methods, our method has a better control of FDR, in particular for genes and transcripts with high inferential uncertainty on abundance.

email: anqizhu@live.unc.edu

11c. A RAPID STEPWISE MAXIMUM LIKELIHOOD PROCEDURE FOR AN ISOLATION-WITH-MIGRATION MODEL

Jieun Park*, Auburn University
Yujin Chung, Kyonggi University, South Korea

An Isolation-with-Migration (IM) model explains the genetic divergence of two populations split away from their common ancestral population. A standard probabilistic model takes account of two levels of uncertainty: (1) the distribution of genetic data given a genealogy and (2) the distribution of a genealogy given an IM model. Under an IM model, a genealogy can contain two kinds of evolutionary paths: vertical inheritance paths through generations and migration paths between populations. The computational complexity of IM model inference is one of the major limitations to analyze genomic data. We propose a rapid stepwise maximum likelihood approach to estimate IM models from genomic data. The first step analyzes genomic data and maximizes the likelihood of a coalescent tree of genealogy containing vertical paths. This first step can algebraically find the maximum likelihood estimation of a coalescent tree. The second step analyzes the estimated trees and finds the parameter values of an IM model which maximizes the distribution of

coalescent tree after taking account of possible migration events. We evaluate the performance of the new method by simulated and real data analyses.

email: jzp0037@auburn.edu

11d. KERNEL ASSOCIATION TEST FOR RARE COPY NUMBER VARIANTS USING PROFILE CURVES

Amanda Brucker*, North Carolina State University
Wenbin Lu, North Carolina State University
Rachel Marceau West, North Carolina State University
Jin Szatkiewicz, University of North Carolina, Chapel Hill
Jung-Ying Tzeng; North Carolina State University

Copy number variants (CNVs) are the gain or loss of DNA segments in the genome that can vary in dosage and length. Rare CNVs have been shown to be associated with phenotypes including cancers and psychiatric disorders. Recent studies have shown that kernel-based tests are a powerful tool for assessing the aggregate effect of rare CNVs on phenotypes due to their ability to capture between- and within-locus etiological heterogeneity. To perform a kernel association test, a locus unit is defined so that similar variants in a single locus correspond to similarity in profiles; then similarity across all loci is summarized in a kernel. In this work, we propose a new kernel-based test that does not require a definition of locus but makes use of the alignment of CNV profiles performed in standard CNV data processing. This test has power to detect associations driven by CNV dosage, length, and dosage-length interactions. In a variety of simulation settings, the proposed test shows comparable and improved power over existing approaches. A real data analysis examines the performance of the test applied to known gene pathways for data from the Psychiatric Genomics Consortium.

email: amanda.brucker@gmail.com

11e. SPARSE NEGATIVE BINOMIAL MODEL-BASED CLUSTERING FOR RNA-Seq COUNT DATA

Md Tanbin Rahman*, University of Pittsburgh
Tianzhou Ma, University of Maryland
George Tseng, University of Pittsburgh

Clustering with variable selection is a challenging but critical task for modern small-n-large-p data. Existing methods based on Gaussian mixture models or sparse K-means provides a solution to continuous data. With the prevalence of RNA-seq technology and lack of count data modeling for clustering, the current practice is to normalize count expression data into continuous measures and apply existing models with Gaussian assumption. In this paper, we develop a negative binomial mixture model with gene regularization to cluster samples (small n) with high-dimensional gene features (large p). EM algorithm and Bayesian information criterion are used for inference and determining tuning parameters. The method is compared with sparse Gaussian mixture model and sparse K-means using extensive simulations and two real transcriptomic applications in breast cancer and rat brain studies. The result shows superior performance of the proposed count data model in clustering accuracy, feature selection and biological interpretation by pathway enrichment analysis.

email: MDR56@pitt.edu

11f. A HIERARCHICAL BAYES MODEL FOR BACKGROUND CORRECTION OF PROTEIN MICROARRAYS

Sophie Berube*, Johns Hopkins University Bloomberg School of Public Health
Thomas A. Louis, Johns Hopkins University Bloomberg School of Public Health

Protein microarrays have the potential to provide information about various biological mechanisms, population level disease and even have the capacity to act as informative biomarkers. However, to extract this information from a protein array, attention must be paid to removing noise while retaining the biologically relevant signal. While methods have been developed to correct for background on DNA microarrays, the variable size of proteins, their relatively weak binding affinity and complex structure warrants methods optimized for the protein mircroarray. To this end, we propose a hierarchical Bayesian model that includes measured foreground and background signals at each probe. Markov Chain Monte Carlo produces the posterior distribution for the true signal; the posterior mean and other summaries are available, or the full distribution can input to the next phase of analysis. We show that this model leaves high intensity signals for which background noise has low impact close to the directly computed estimate. Low intensity signals receive a substantial and beneficial adjustment, while keeping the mass of the posterior distribution on positive values.

email: sberube3@jhmi.edu

11g. STATISTICAL INFERENCE OF HIGH-DIMENSIONAL MODIFIED POISSON-TYPE GRAPHICAL MODELS WITH APPLICATION TO CHILDHOOD ASTHMA IN PUERTO RICANS

Rong Zhang*, University of Pittsburgh
Zhao Ren, University of Pittsburgh
Wei Chen, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh

Recent advances in next-generation sequencing (NGS) have yielded a large amount of data from different living systems such as DNA, RNA and proteomics. The discreteness and the high dimensions of these NGS data have posed great challenges in biological network analysis. Although estimation theories for four high-dimensional modified Poisson-type graphical models have been proposed to tailor the network analysis of those count-valued data, the rigorous statistical inference of these models is still largely unknown. We herein propose a novel two-step procedure in statistical inference of each edge for these modified Poisson-type graphical models using a cutting-edge generalized low-dimensional projection approach for bias correction. Unlike the existing method to continuize the discrete values for further analysis, which to some extent removes the intrinsically useful information in data, our method maintains the biological meanings of those count values. A numerical simulation study illustrates more accurate inferential results compared to the conventional ones. We have also applied our method to a novel RNA-seq gene expression data in childhood asthma in Puerto Ricans.

email: roz16@pitt.edu

11h. SPARSE SEMIPARAMETRIC CANONICAL CORRELATION ANALYSIS FOR DATA OF MIXED TYPES

Grace Yoon*, Texas A&M University
Raymond J. Carroll, Texas A&M University
Irina Gaynanova, Texas A&M University

Canonical correlation analysis investigates linear relationships between two sets of variables, but often works poorly on modern data sets due to high-dimensionality and mixed data types (continuous/binary/zero-inflated). We propose a new approach for sparse canonical correlation analysis of mixed data types that does not require explicit parametric assumptions. Our main contribution is the use of truncated latent Gaussian copula to model the data with excess zeroes, which allows us to derive a rank-based estimator of latent correlation matrix without the estimation of marginal transformation functions. The resulting semiparametric sparse canonical correlation analysis method works well in high-dimensional settings as demonstrated via numerical studies, and application to the analysis of association between gene expression and micro RNA data of breast cancer patients.

email: gyoon@stat.tamu.edu

12. POSTERS: FUNCTIONAL DATA/HIGH DIMENSIONAL

12a. COMPUTATIONAL METHODS FOR DYNAMIC PREDICTION

Andrada E. Ivanescu*, Montclair State University
William Checkley, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University

We discuss algorithmic and software methods for a class of dynamic regression models designed to predict the future of growth curves based on their historical dynamics. This class of models incorporates both baseline and time-dependent covariates, start with simple regression models and build up to dynamic function-on-function regressions. We focus on software implementations of the methods and we use R statistical software.

email: ivanescua@montclair.edu

12b. MODELING CONTINUOUS GLUCOSE MONITORING (CGM) DATA DURING SLEEP

Irina Gaynanova*, Texas A&M University
Naresh M. Punjabi, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University

We introduce a multilevel functional Beta model to quantify the blood glucose levels measured with continuous glucose monitors (CGMs) for multiple days in study participants with Type 2 diabetes mellitus. We focus on the actigraphy-estimated sleep periods to reduce the effects of meals and physical activity. The model produces confidence intervals for blood glucose levels at any time from the actigraphy-estimated sleep onset, quantifies the within- and between-subject variability of blood glucose, and produces interpretable parameters of the blood dynamics during the actigraphy-estimated sleep periods. We provide visualization tools for our results and validate the estimated model parameters versus levels of Hemoglobin A1c.

email: irinag@stat.tamu.edu

12c. SAMPLING STUDIES FOR LONGITUDINAL FUNCTIONAL DATA ANALYSIS

Toni L. Jassel*, Montclair State University
Andrada Ivanescu, Montclair State University

We study the data setting consisting of functional data samples repeatedly observed over time. The focus is on the dynamic prediction of the future trajectory. Regression methods based on dynamic functional regression are used for prediction. We propose strategies for the selection of the sampling design for longitudinal functional data. An application to child growth is presented.

email: JasselT1@montclair.edu

12d. WEARABLE DEVICES ARE OBJECTIVE BUT IMPERFECT - TOWARDS CORRECTING FOR TWO SOURCES OF ERROR

Dane R. Van Domelen*, Johns Hopkins University
Vadim Zipunnikov, Johns Hopkins University

Wearable devices are increasingly used to measure physical activity and sleep in epidemiological and clinical studies. While sensor-based data are objective, signal-derived measures are susceptible to two sources of error that can compromise validity: device-related error (e.g. recorded steps ≠ true steps) and sampling-related error (e.g. 7-day average ≠ long-term average). We use accelerometry data from the National Health and Nutrition Examination Survey to illustrate how the standard practice of analyzing accelerometer data tends to attenuate associations with health outcomes. We note that bias due to both errors would disappear with a longer wear protocol, assuming purely random errors, but most studies use 7 days. For data like NHANES, with no "gold standard" accompanying the accelerometer data, corrective methods will typically require treating daily measurements as replicates. We propose a likelihood-based approach that addresses the challenges above by assuming (1) both errors are mean-one lognormal and multiplicative; and (2) subject-specific long-term weekday and weekend averages (latent variables) are bivariate lognormal conditional on covariates.

email: vandomed@gmail.com

13. POSTERS: SPATIAL/TEMPORAL MODELING

13a. SPATIAL STATISTICAL METHODS TO ASSESS THE RELATIONSHIP BETWEEN WATER VIOLATIONS AND POVERTY AT THE COUNTY LEVEL: IN AMERICA, WHO HAS ACCESS TO CLEAN WATER?

Ruby Lee Bayliss*, Drexel University
Loni Phillip Tabb, Drexel University

The Flint, Michigan water crisis that occurred in 2014, has brought national attention to the importance of safe drinking water. The existence of water violations within

community water systems since 2014 has been thoroughly documented throughout the United States, but the relationship between people who are socioeconomically disadvantaged and the occurrence of water violations is unclear. The objective of this study is to determine if there is a relationship between people with low socioeconomic status and access to clean community water systems on the state and county level using county level data found in the County Health Rankings database supported by the Robert Wood Johnson Foundation. We identified three states that represented the smallest, moderate, and greatest proportions of counties that had at least one community water violation across the United States. Utilizing both exploratory and inferential spatial statistical methods, we examined the spatial patterning of the community water violations as well as the neighborhood characteristics within each state, with the goal to further understand this relationship.

email: rlb96@drexel.edu

13b. DESCRIBING THE SPATIOTEMPORAL PATTERNING OF OVERALL HEALTH IN THE UNITED STATES USING COUNTY HEALTH RANKINGS FROM 2010-2018

Angel Gabriel Ortiz*, Drexel University
Loni Tabb, Drexel University

Overall county health rankings have been obtained annually since 2010 by the University of Wisconsin and the Robert Wood Johnson Foundation. These measures are helpful in identifying the spatiotemporal health implications and the driving factors that may explain health disparities at the county level. This study will describe the spatiotemporal patterning of overall health across the country between the years 2010-2018. Health outcomes such as length of life and quality of life and their associations with county level demographic measures will be used in creating Bayesian models. The Integrated Nested Laplace Approximation (R-INLA) package is used to obtain approximation of the parameters of interest. The study is currently on-going however; previous studies have shown that health outcomes present spatial clustering that have important implications on overall health (Tabb, McClure, Quick, Purtle, & Roux, 2018). This study will extend previous findings by incorporating a temporal factor to further observe how spatial trends compare for the years 2010-2018.

email: ago27@drexel.edu

13c. DISCRIMINANT ANALYSIS FOR LONGITUDINAL MRI

Rejaul Karim*, Michigan State University
Taps Maiti, Michigan State University
Chae Young Lim, Seoul National University

Linear Discriminant Analysis is one of the popular classification methods. Main focus is sparse representation of Fischer discriminant direction. The asymptotic properties of this classifier is investigated under spatio-temporal paradigm. The results are verified on high dimensional data sets.

email: karimrej@stt.msu.edu

13d. PROPENSITY SCORE MATCHING FOR MULTI-LEVEL AND SPATIAL DATA

Behzad Kianian*, Emory University
Howard H. Chang, Emory University
Rachel E. Patzer, Emory University
Lance A. Waller, Emory University

In many health studies, an individual's treatment assignment and health outcomes are determined both by where they live and where they seek care. We consider propensity score matching in analyzing multi-level observational data with a binary treatment in the presence of unmeasured cluster-level and spatial confounders. We present a systematic simulation study comparing various existing approaches, such as propensity score models that include spatial coordinates or cluster-specific random/fixed intercepts, and a recently developed method of distance-adjusted propensity score matching. We additionally propose a method that synthesizes two methods recently developed in a two-stage procedure: (1) match within the same cluster where possible; (2) for the remaining treated units that were unmatched in the first stage, require that remaining matches are spatially close. The method is applied to a study of kidney disease patients who have recently started dialysis. Patients are reported as either informed of their transplant options or not; this decision and the patient's outcomes are likely impacted by individual, facility, and area-level factors.

email: behzad.kianian@gmail.com

13e. MULTIPLE TESTING AND ESTIMATION OF DISEASE ASSOCIATIONS BASED ON SEMI-PARAMETRIC HIERARCHICAL MIXTURE MODELS, POSSIBLY INCORPORATING BRAIN AREAS

Ryo Emoto*, Nagoya University School of Medicine
Takahiro Otani, Nagoya University School of Medicine
Shigeyuki Matsui, Nagoya University School of Medicine

In disease-association studies using neuroimaging data, both the detection of disease-associated areas of the brain and the estimation of the magnitudes of the associations or effect sizes for individual brain areas are performed. In this paper, as a modelling of whole-brain associations, we propose to use a semi-parametric hierarchical mixture model with a hidden Markov random field structure to incorporate the spatial dependency among contiguous voxels and a non-parametric effect size distribution to flexibly estimate the underlying effect size distribution. Based on this model, we derive procedures for multiple testing and estimation of disease associations for individual voxels or brain areas. An application to neuroimaging data from an Alzheimer's disease study is provided.

email: emoto.ryo@b.mbox.nagoya-u.ac.jp

13f. TRENDS IN TRACT-LEVEL DENTAL VISIT RATES IN PHILADELPHIA BY RACE, SPACE AND TIME

Guangzi Song*, Drexel University
Harrison S. Quick, Drexel University

Oral health is an essential and integral component of an individual's overall health. Delaying dental visit is indicated to be strongly correlated with poor oral health and overall health. The factors influencing delayed dental visits includes general and

regional socioeconomic and demographic conditions, individual behaviors, and oral health policies, etc. The objective of this work is to investigate the disparities of race/age/sex-specific regular dental visit rates by poverty status in Philadelphia census tracts over the period 2000-2015. To analyze and obtain more reliable rate estimates, we apply a multivariate spatiotemporal Bayesian model, simultaneously accounting for spatial-, temporal-, and between-race dependence structures. The results suggest that regular dental visit rates have changed at the city-level for men and women of all three races studied, the magnitude and geographic distribution of these changes differ by race and sex. In conclusion, the model used here has expanded our knowledge of dental visit rates in Philadelphia. We believe this approach can be used to explore trends in a variety of health measures from across the nation.

email: gs556@drexel.edu

14. POSTERS: PUBLIC HEALTH/SURVEYS/EHR

14a. SUPERVISED DIMENSION REDUCTION USING BAYESIAN HIERARCHICAL MODELING: A SIMULATION STUDY AND APPLICATION TO AMBIENT AIR POLLUTANTS

Ray Boaz*, Medical University of South Carolina
Andrew Lawson, Medical University of South Carolina
John Pearce, Medical University of South Carolina

Risk associated with air pollution has typically been evaluated at an individual pollutant level. Researchers understand that the presence of multiple pollutants may have interactive and grouped effects not currently captured in epidemiologic research. We have developed a novel mixture classification and modeling technique to characterize air pollution exposure in a more realistic context that accounts for the simultaneous and joint nature of the exposure. The model uses a method that informs the grouping of mixtures based on the health outcome of interest within a Bayesian Hierarchical Modeling framework. We have conducted simulation studies with ground truth scenarios consisting of mixtures of pollutants X impacting an outcome Y. The pollutants X have prespecified groupings with deterministic impact on the outcome Y, so the model parameters have been evaluated for fidelity to the prescribed relationship. We have also evaluated our model's accuracy and impact using previously developed simulation data sets from NIEHS. Our model has successfully identified groupings of variables and qualitative effects to this point.

email: boaz@musc.edu

14b. COMPARISON OF INTERVAL ESTIMATION METHODS FOR DOSE-RESPONSE RELATIONSHIP: FREQUENTIST MODEL AVERAGING (FMA) VERSUS CORRECTED CONFIDENCE INTERVAL ESTIMATION (CCI) WHEN EXPOSURE UNCERTAINTY IS COMPLEX

Deukwoo Kwon*, University of Miami

Risk estimation from dose-response analysis for some epidemiologic studies can be difficult task since exposure uncertainty is complex. In this situation, single exposure measurement per subject given from exposure assessment experts does not take account of complex exposure uncertainty. Simon et al. (2015) provided advanced exposure assessment technique to consider complex uncertainty using the two-dimensional Monte Carlo (2DMC) method. From the 2DMC method, multiple exposure realizations per subject can be provided for dose-response analysis in epidemiological studies. To date, several approaches to dose-response analysis have been published that address situations with multiple exposure realizations: Stayner et al (2007), Kwon et al. (2016) and Zhang et al. (2017). In this study, we proposed a frequentist model averaging (FMA) approach and show that FMA and BMA have comparable performance and then investigate whether FMA and CCI work well in various scenarios in terms of exposure complexity using simulation study and real data example.

email: DKwon@med.miami.edu

14c. JOINT MODELLING OF BINARY AND CONTINUOUS MEASUREMENTS IN LARGE HEALTH SURVEYS AND ITS APPLICATION TO NETWORK ANALYSIS, FRAILTY, AND MORTALITY IN NHANES 1999-2010

Debangan Dey*, Johns Hopkins Bloomberg School of Public Health
Irina Gaynanova, Texas A&M University
Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health

Network analysis has rapidly gained popularity in neuroimaging, genomics and other scientific domains. However, a little has been done to adapt network analysis to heterogeneous measurements collected by national health surveys and biobanks. This is primarily due to a lack of understanding on how to jointly model multiple comorbidities, health deficits, and health biomarkers, often recorded via binary and continuous measurements. Our approach adapts a recently proposed semiparametric gaussian copula that estimates a latent correlation structure of mixed type (binary and continuous) random vectors. After estimating joint distribution of latent continuous variables, we build a network by applying GLASSO to control for number of connections. The method provides a flexible framework which allows to jointly model outcome and predictors, impute missing data, perform dimension reduction on the latent space and do prediction on the latent and observed space. We demonstrate this method on 47 binary and continuous variables typically included in Frailty Index (FI). Using latent principal components and network connectivity's, a few weighted versions of FI are developed and compared.

email: ddey1@jhu.edu

14d. A METHODOLOGY AND SAS MACRO TO ESTIMATE CONSUMPTION OF ALCOHOL FROM SURVEY DESIGNS: APPLICATION TO NHANES AND NLSY

Elysia A. Garcia*, University of Texas Health Science Center
Stacia M. DeSantis, University of Texas Health Science Center

It is often of interest to measure substance intake from national survey data in order to perform risk assessments and guide public health recommendations. Current methods for estimating average intake for count outcomes do not account for the complex survey design elements. We introduce a method for estimating population mean intake, distributions, and standard errors for substances measured as counts. A negative binomial hurdle is used to estimate the product of the probability of consuming and the amount consumed over multiple repeated observations, allowing for correlation between the two parts. Using these model estimates, distributions of intake are then simulated to calculate population mean intake. Standard errors are estimated using balanced repeated replication method, which accounts for

strata, clusters, and weights. We illustrate the utility of the method by estimating alcohol intake from a cross-sectional and longitudinal survey - the National Health and Nutritional Examination Survey and the National Longitudinal Survey of Youth, respectively. These applications allow us to estimate mean and lifetime mean intake, respectively, by subgroups of interest.

email: Elysia.A.Garcia@uth.tmc.edu

14e. A PRIVACY-PRESERVING AND COMMUNICATION EFFICIENT DISTRIBUTED ALGORITHM FOR LOGISTIC REGRESSION WITH EXTREMELY RARE OUTCOMES OR EXPOSURES

Rui Duan*, University of Pennsylvania
Mary Regina Boland, University of Pennsylvania
Jason H. Moore, University of Pennsylvania
Yong Chen, University of Pennsylvania

Electronic Health Records (EHR) contain extensive information on various health outcomes and risk factors, and therefore have been broadly used in healthcare research. Institutional data integration is a major trend in EHR-based research. In particular, for studying relatively rare events or conditions, such as complications from invasive procedures, adverse events associated with new medications, association of disease with a rare gene variant, and many others, integrating EHR data from different clinical sites is critical for obtaining more accurate, generalizable and reproducible results. To overcome the barrier of patient-level data sharing, we propose a privacy-preserving and communication-efficient distributed algorithm for logistic regression. Our simulation study showed that our method is powerful for studying rare outcomes or rate exposures compared to many existing methods. In most of the cases, our algorithm reached comparative accuracy comparing to the oracle estimator where data are pooled together. We applied our algorithm to an EHR data from the University of Pennsylvania health system to evaluate the risks of fetal loss due to various medication exposures.

email: duanrui1991@gmail.com

14f. BAYESIAN METHODS FOR ESTIMATING THE POPULATION ATTRIBUTABLE RISK IN THE PRESENCE OF EXPOSURE MISCLASSIFICATION

Benedict Wong*, Harvard T.H. Chan School of Public Health
Molin Wang, Harvard T.H. Chan School of Public Health
Lorenzo Trippa, Harvard T.H. Chan School of Public Health
Donna Spiegelman, Yale School of Public Health

Estimation of the population attributable risk (PAR) has become an important goal in public health research, because they describe the proportion of disease cases that could be prevented if a set of exposures were eliminated from a target population as a result of some intervention. In epidemiologic studies, binary and categorical covariates are often misclassified, leading to bias in PAR estimates. We present both approximate and full Bayesian methods for point and interval estimation for the PAR in the presence of misclassification. We consider multiple scenarios, including both main study/internal validation study and main study/external validation study designs, with and without the assumption of transportability of the exposure prevalence's. We apply these methods to estimate the full and partial PARs in the Health Professionals Follow-Up Study of dietary risk factors for colorectal cancer.

email: benwong2016@gmail.com

14g. EVALUATING THE EFFECTS OF ATTITUDES ON HEALTH-SEEKING BEHAVIOR AMONG A NETWORK OF PEOPLE WHO INJECT DRUGS

Ashley Buchanan*, University of Rhode Island
Ayako Shimada, Thomas Jefferson University
Natallia Katenka, University of Rhode Island
Samuel Friedman, National Development and Research Institutes, Inc.

Like many other populations, people who inject drugs (PWIDs) are embedded in social networks or communities (e.g., injection drug, non-injection drug, sexual risk network) and exert biological and social influence on the members of their networks. The direct effect is the effect on the participants who were exposed and the disseminated effect is the effect on the participants who shared a network with the exposed participant. We analyzed a network of PWIDs from the Social Factors and HIV Risk Study (SFHR), conducted between 1991 and 1993 in Bushwick, New York, where links were defined by shared sexual and injection behaviors. After identifying communities in the SFHR network with a modularity-based community detection, a network-based causal inference methodology for clustered observational data was employed to assess the effect of attitudes toward HIV/AIDS risk among PWIDs on their own health-seeking behavior and that of other members in their community. In addition, we evaluated the impact of community structure on likelihood of the outcomes. Our results provide evidence of disseminated effects of attitudes among PWIDs networks.

email: buchanan@uri.edu

15. NOVEL NEUROIMAGING METHODS FROM PROCESSING TO ANALYSIS

A UNIFIED FRAMEWORK FOR BRAIN FUNCTIONAL CONNECTIVITY USING COVARIANCE REGRESSION

Ani Eloyan*, Brown University

Functional connectivity (FC) has been extensively used in the imaging literature to study functional associations among brain regions and identify temporal correlations between neurophysiological events. FC often refers to the identification of undirected temporal associations between any two regions in the brain, often including spatially incongruous areas. While FC has been widely studied by researchers in the area, no unified modeling framework exists for estimating FC or brain networks and conducting further inference comparing populations. In this talk, I will review various approaches to FC estimation and inference and present a novel general statistical framework for FC analysis in populations. The proposed framework incorporates the special structure of the connectivity maps using the covariance regression model to draw inferences of model parameters under the correct modeling assumptions for comparing populations or identifying changes of connectivity over time. I will present comparisons of visual-motor connectivity at rest

between people with Autism spectrum disorders and their neurotypical peers using the publicly available ABIDE database.

email: ani_eloyan@brown.edu

ROBUST SPATIAL EXTENT INFERENCE WITH A SEMIPARAMETRIC BOOTSTRAP JOINT TESTING PROCEDURE

Simon Vandekar*, Vanderbilt University
Theodore D. Satterthwaite, University of Pennsylvania
Cedric H. Xia, University of Pennsylvania
Kosha Ruparel, University of Pennsylvania
Ruben C. Gur, University of Pennsylvania
Raquel E. Gur, University of Pennsylvania
Russell T. Shionhara, University of Pennsylvania

Spatial extent inference (SEI) is widely used across neuroimaging modalities to study brain-behavior associations that inform our understanding of disease. Recent studies have shown that Gaussian random field (GRF) based tools can have inflated family-wise error rates (FWERs). The failure of GRF-based methods is due to unrealistic assumptions about the covariance function of the imaging data. The permutation procedure is the most robust SEI tool because it estimates the covariance function from the imaging data. However, the permutation procedure can fail because exchangeability is violated in many imaging modalities. Here, we propose a semiparametric bootstrap joint (sPBJ) testing procedures that are designed for SEI of multilevel imaging data. The sPBJ uses a robust estimate of the covariance function, which yields consistent estimates of standard errors, even under model misspecification. We use our methods to study the association between performance and executive functioning in a working memory fMRI study. The sPBJ has equal or superior power to the other procedures while maintaining the nominal type 1 error rate.

email: simon.vandekar@vanderbilt.edu

ADDRESSING PARTIAL VOLUME EFFECTS USING INTRA-SUBJECT LOCALLY ADJUSTED CEREBRAL BLOOD FLOW IMAGES

Kristin A. Linn*, University of Pennsylvania
Simon Vandekar, Vanderbilt University
Russell T. Shinohara, University of Pennsylvania

Local cortical coupling is a subject-specific measure of the spatially varying relationship between cortical thickness and sulcal depth. Although it is a promising first step towards understanding local covariance patterns between two image-derived measurements, a more general coupling framework that can accommodate multiple volumetric imaging modalities is warranted. We first introduce Inter-Modal Coupling (IMCo), an analogue of local coupling in volumetric space that can be used to produce subject-level, spatially varying feature maps derived from two volumetric imaging modalities. We then leverage IMCo to address partial volume effects when studying localized relationships between gray matter density and cerebral blood flow (CBF) among participants in the Philadelphia Neurodevelopmental Cohort. We show that when CBF images are adjusted for partial volume effects at the subject level using our method, we have more power to detect non-linear interactions between age and sex in voxelwise analyses. We call the proposed IMCo-adjusted CBF images Intra-Subject Locally Adjusted Cerebral Blood Flow (ISLA-CBF) images.

email: klinn@pennmedicine.upenn.edu

16. STATISTICAL CHALLENGES AND OPPORTUNITIES FOR ANALYSIS OF LARGE-SCALE OMICS DATA

LINEAR HYPOTHESIS TESTING FOR HIGH DIMENSIONAL GENERALIZED LINEAR MODELS

Runze Li*, The Pennsylvania State University
Chengchun Shi, North Carolina State University
Rui Song, North Carolina State University
Zhao Chen, Fudan University

This paper is concerned with testing linear hypotheses in high-dimensional generalized linear models. To deal with linear hypotheses, we first propose constrained partial regularization method and study its statistical properties. We further introduce an algorithm for solving regularization problems with folded-concave penalty functions and linear constraints. To test linear hypotheses, we propose a partial penalized likelihood ratio test, a partial penalized score test and a partial penalized Wald test. We show that the limiting null distributions of these three test statistics are chi-square distribution with the same degrees of freedom, and under local alternatives, they asymptotically follow non-central chi-square distributions with the same degrees of freedom and noncentral parameter, provided the number of parameters involved in the test hypothesis grows to infinity at a certain rate. Simulation studies are conducted to examine the finite sample performance of the proposed tests. Empirical analysis of a real data example is used to illustrate the proposed testing procedures.

email: rzli@psu.edu

SCALABLE WHOLE GENOME SEQUENCING ASSOCIATION ANALYSIS USING FUNCTIONAL ANNOTATION AND CLOUD COMPUTING

Xihong Lin*, Harvard University

Whole-genome sequencing (WGS) studies have been increasingly conducted to investigate associations between susceptible rare variants and disease/traits. The existing variant-set tests only incorporate the minor allele frequency (MAF) as weights in rare variant association analysis. External biological information based on various functional annotations can be utilized to predict variant functionality, and thus help boost power for variant-set tests. We develop Genome Analysis Program and Platform (GAPP) for scalable analysis of WGS data. GAPP include several scalable new methods our group has developed, including variant-Set Test for Association using Annotation infoRmation (STAAR), a general framework that incorporates multiple complementary functional annotations as an omnibus weighting scheme to boost power for variant-set tests in WGS studies; SCANG, which is a scan procedure for detecting signal regions in WGS association studies. We will discuss the use of cloud

for analyzing WGS data. Applications to the WGS data from the Genome Sequencing Program (GSP) of NHGRI and the TOPMed of NHLBI will be discussed.

email: xlin@hsph.harvard.edu

LEARNING HIERARCHICAL INTERACTIONS IN HIGH DIMENSIONS

Lingzhou Xue*, The Pennsylvania State University

In this talk, we will introduce an innovative nonconvex learning of hierarchical interactions in high-dimensional statistical models. We first use the affine sparsity constraints to provide a precise characterization of both strong and weak hierarchical interactions. However, these affine sparsity constraints do not lead to a closed feasible region. To address this issue, we derive the explicit closure of the affine sparsity constraint for learning nonconvex hierarchical interactions. We prove that the desired oracle-like solution can be found by solving a sequence of folded concave penalized estimation and the desired strong or weak hierarchy holds with probability one. Furthermore, we study the asymptotic properties for learning hierarchical interactions using the folded concave penalized estimation. We will demonstrate the power of our proposed methods in biological applications.

email: lzxue@psu.edu

ANALYSIS OF IMAGING GENETIC "BIG DATA SQUARED" STUDIES

Heping Zhang*, Yale University School of Public Health

Neuroimaging has been an essential tool for collecting data on the functioning of brain. High throughput technologies have provided ultra-dense genetic markers to enable us in identifying genetic variants for complex diseases. Only until recently, datasets of reasonably large scale become available that contain both imaging and genetic data. Due to the complexities and high dimensionality in such data, most of the existing datasets are still relatively small in sample sizes but larger datasets are in the horizon. Thus, it is timely and important to develop statistical methods and analytic tools to analyze imaging genetic data – the so-called big data squared. In this talk, I will present the basic technologies, concepts, challenges, and methods related to imaging genetic data. I will use specific data on learning disorders illustrate how to quantify neurobiological risk for learning problems with neuroimaging biomarkers and how to integrate imaging and genetic data in our understanding of cognition and genetic etiologies. This is a joint work with Hongtu Zhu, Xuan Bi, Long Feng, Canhong Wen, and Chintan Mehta. This research is supported in part by NIH grants.

email: heping.zhang@yale.edu

17. LONGITUDINAL AND FUNCTIONAL MODELS FOR PREDICTING CLINICAL OUTCOMES

BAYESIAN REGRESSION MODELS FOR BIG SPATIALLY OR LONGITUDINALLY CORRELATED FUNCTIONAL DATA

Jeffrey S. Morris*, University of Texas MD Anderson Cancer Center
Hongxiao Zhu, Virginia Tech University
Hojin Yang, University of Texas MD Anderson Cancer Center
Michelle Miranda, University of Texas MD Anderson Cancer Center
Wonyul Lee, U.S. Food and Drug Administration
Veera Baladandayuthapani, University of Michigan
Birgir Hrafnkelsson, University of Iceland

A series of Bayesian regression modeling strategies that can be used for high-dimensional spatially- or longitudinally correlated functional data will be described. Intrafunctional correlation is handled through basis function modeling, while interfunctional correlation is captured by one of three approaches: (1) parametric or nonparametric random effect functions, (2) separable or non-separable spatial (or temporal) inter-functional processes, or (3) tensor kernels.. Rigorous Bayesian inference is done in such a way that accounts for any potential multiple testing issues. We will describe these general approaches and illustrate them on a series of complex, high-dimensional, spatially and longitudinally correlated functional data sets coming from strain tensor data from a glaucoma study, event-related potential data from a smoking cessation study, and characterizing climate change in Icleand over the past 7 decades.

email: jefmorris@mdanderson.org

JOINT MODELING OF MULTIVARIATE LONGITUDINAL DATA WITH A BINARY RESPONSE

Paul S. Albert*, National Cancer Institute, National Institutes of Health
Sung Duk Kim, National Cancer Institute, National Institutes of Health

Predicting binary events such as newborns with large birthweight is important for obstetricians in their attempt to reduce both maternal and fetal morbidity and mortality. Such predictions have been a challenge in obstetric practice, where longitudinal ultrasound measurements taken at multiple gestational times during pregnancy may be useful for predicting various poor pregnancy outcomes. We develop a flexible class of joint models for the multivariate longitudinal ultrasound measurements that can be used for predicting a binary event at birth. A skewed multivariate random effects model is proposed for the ultrasound measurements, and the skewed generalized t-link is assumed for the link function relating the binary event and the underlying longitudinal processes (Kim and Albert, Biometrics 2016). We consider a shared random effect to link the two processes together. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. The proposed methodology is illustrated with data from the NICHD Successive Small-for-Gestational-Age Births study, a large prospective fetal growth cohort conducted in Norway and Sweden.

email: albertp@mail.nih.gov

MODELING OF HIGH-DIMENSIONAL CLINICAL LONGITUDINAL OXYGENATION DATA

Abdus Sattar*, Case Western Reserve University
Seunghee Margevicius, Case Western Reserve University

Motivated by oxygenation of retinopathy of prematurity (ROP) study, we developed a penalized spline mixed effects model for a high-dimensional nonlinear longitudinal continuous response variable using the Bayesian approach. The ROP study is

complicated by the fact that there are non-ignorable missing response values. To address the non-ignorable missing data in the Bayesian penalized spline model, we applied a selection model. Properties of the estimators are studied using Markov Chain Monte Carlo (MCMC) simulation. The proposed new approach joint model did better compare to the semiparametric mixed effects model with non-ignorable missing values in terms of bias and percent bias. We performed sensitivity analysis for the hyper-prior distribution choices for the variance parameters of spline coefficients on the proposed joint model. We also applied our novel method to the sample entropy data in ROP study for handling nonlinearity and the non-ignorable missing response variable.

email: sattar@case.edu

18. RECENT BAYESIAN METHODS FOR CAUSAL INFERENCE

ASSESSING CAUSAL EFFECTS IN THE PRESENCE OF TREATMENT SWITCHING THROUGH PRINCIPAL STRATIFICATION

Fabrizia Mealli*, University of Florence

Clinical trials, focusing on survival outcomes for patients suffering from AIDS and painful cancers, often allows patients to switch arm if their physical conditions are worse than certain tolerance levels. The ITT analysis does not give information about the effect of the actual treatment receipt, and other methods propose to reconstruct the outcome a unit would have had if s/he had not switched and rely on strong assumptions. We propose to re-define the problem of treatment switching using principal stratification; and introduce new causal estimands. We use a Bayesian approach to account for (i) switching time happening in continuous time generating a continuum of principal strata; (ii) switching time not being defined for units who never switch; and (iii) both survival time and switching time are subject to censoring. We illustrate our framework using simulated data based on the Concorde study, a randomized controlled trial to assess the effect on time-to-disease progression or death of immediate versus deferred treatment with zidovudine among patients with asymptomatic HIV infection.

email: fabrizia.mealli@unifi.it

BAYESIAN NONPARAMETRIC MODELS WITH FASTER ALGORITHMS FOR ESTIMATING CAUSAL EFFECTS

Jason Roy*, Rutgers University

Bayesian nonparametric (BNP) models are useful for causal inference from observational studies, because they allow for adjustment of confounders without making strong parametric assumptions. However, many algorithms for fitting BNP models are computationally expensive. We consider here several ways of speeding up the computations. Results are compared via simulations studies.

email: jason.roy@rutgers.edu

RECIPROCAL GRAPHICAL MODELS FOR INTEGRATIVE GENE REGULATORY NETWORK ANALYSIS

Peter Mueller*, University of Texas, Austin
Yang Ni, Texas A&M University
Yuan Ji, University of Chicago

Constructing gene regulatory networks is a fundamental task in systems biology. We introduce a Gaussian reciprocal graphical model for inference about gene regulatory relationships by integrating mRNA gene expression and DNA level information including copy number and methylation. Data integration allows for inference on the directionality of certain regulatory relationships, which would be otherwise indistinguishable due to Markov equivalence. Efficient inference is developed based on simultaneous equation models. Bayesian model selection techniques are adopted to estimate the graph structure. We illustrate our approach by simulations and application in colon adenocarcinoma pathway analysis.

email: pmueller@math.utexas.edu

A BAYESIAN SEMIPARAMETRIC FRAMEWORK FOR CAUSAL INFERENCE IN HIGH-DIMENSIONAL DATA

Joseph Antonelli*, University of Florida
Francesca Dominici, Harvard T.H. Chan School of Public Health

We introduce a Bayesian framework for estimating causal effects of binary and continuous treatments in high-dimensional data. The proposed framework extends many of the existing semiparametric estimators introduced in the causal inference literature to high-dimensional settings. Our approach has the following features: it 1) considers semiparametric estimators that reduce model dependence; 2) introduces flexible Bayesian priors for dimension reduction of the covariate space that accommodate nonlinearity; 3) provides posterior distributions of any causal estimator that can be defined as a function of the treatment and outcome model; 4) provides posterior credible intervals with improved finite sample coverage compared to frequentist measures of uncertainty which rely on asymptotic properties. We show that the posterior contraction rate of the proposed doubly robust estimator is the product of the posterior contraction rates of the treatment and outcome models, allowing for faster posterior contraction. We illustrate our approach via simulation and apply it to estimate the causal effect of continuous environmental exposures.

email: jantonelli111@gmail.com

19. NEW METHODS FOR COST-EFFECTIVENESS ANALYSIS IN HEALTH POLICY RESEARCH

COST AND COST-EFFECTIVENESS ANALYSIS: WHERE ARE WE NOW?

Heejung Bang*, University of California, Davis

Effectiveness and cost are the two key components when comparing healthcare options and making treatment decisions (e.g., comparative effectiveness research). Cost-effectiveness analysis (CEA) is the methodological approach that combines

effectiveness and cost, and aids this evaluation. In this talk, we will review the history of cost and cost-effectiveness research along with the basic concepts and statistical methods, and discuss quantitative and qualitative issues, which are quite unique to CEA. Also, cost data are increasingly collected, available and utilized in the big data and machine learning era, so we will discuss data and informatics resources, advances and adaptations in statistical methodologies, and impactful applications for old problems in this new environment.

email: heejungbang@msn.com

AGENT-BASED MODELLING FOR BETTER UNDERSTANDING HEALTH DISPARITIES

Efrén Cruz Cortés*, Colorado School of Public Health
Debashis Ghosh, Colorado School of Public Health

In this talk, we will introduce a class of stochastic models termed agent-based models. We show the effectiveness of these models for better understanding health disparities when there are sensitive populations present. Our consideration of these models is motivated by recent published studies showing how the use of machine learning algorithms have led to a form of discrimination against African-American populations in terms of predicting recidivism. Using very simple agent-based models, we show how simple rules and agents can lead to powerful findings when it comes to understanding disparities between populations. Time permitting, we will also discuss some of the associated causal inference modelling and show how these models can reduce some of the classical assumptions typically used.

email: efren.cruzcortes@ucdenver.edu

APPROACHES TO COST-EFFECTIVENESS ANALYSIS BASED ON SUBJECT-SPECIFIC MONETARY VALUE

Andrew J. Spieker*, Vanderbilt University Medical Center
Nicholas Illenberger, University of Pennsylvania Perelman School of Medicine
Jason A. Roy, Rutgers School of Public Health
Nandita Mitra, University of Pennsylvania Perelman School of Medicine

The net monetary benefit has been used as an aggregate measure of cost and clinical efficacy to inform policy decisions. Briefly, this approach attempts to summarize the extent to which, on average, a treatment's level of efficacy (relative to some control) justifies the cost associated with it at some particular willingness-to-pay threshold. This approach does not provide insights into the proportion of patients that can expect to benefit from the experimental treatment, though such insights could help paint a more complete picture of a treatment's overall advantages and limitations. In this talk, we present the causal net monetary benefit and the cost-effectiveness determination curve to characterize population-level cost-effectiveness and quantify the proportion of pairs of individuals (one from each treatment group) for whom the treatment is deemed cost-effective. We focus on methods to account for confounding, treatment effect heterogeneity, and censoring of costs; and provide illustrations from an analysis of endometrial cancer patients from a SEER-Medicare linked database.

email: andrew.j.spieker@vumc.org

MICROSIMULATIONS FOR COST-EFFECTIVENESS ANALYSIS: MODELING THERAPY SEQUENCE IN ADVANCED CANCER

Elizabeth A. Handorf*, Fox Chase Cancer Center
Andres Correa, Cooper University Health Care
Chethan Ramamurthy, University of Texas Health Science Center at San Antonio
Daniel Geynisman, Fox Chase Cancer Center
J. Robert Beck, Fox Chase Cancer Center

When studying health-economic outcomes, particularly for new treatments, it is common for subject-level cost and effectiveness data to be unavailable. Researchers have therefore developed techniques to synthesize data from published literature. One popular method is microsimulations, where a model framework consisting of health states is developed, and simulated patients move through the model, transitioning from state to state probabilistically. Motivated by a study of metastatic prostate cancer, we develop a general microsimulation framework to estimate the cost-effectiveness of therapy order (A→B vs. B→A). We discuss the advantages of microsimulation models over Markov cohort models, and develop strategies to infer time-dependent state-transition probabilities based on published survival curves. We apply this method to a cost-effectiveness analysis of metastatic prostate cancer, where docetaxel, an older and less expensive chemotherapeutic agent, and abiraterone acetate, a newer and more expensive hormonal therapy, are both approved for first and second line treatment. We demonstrate that the cost-effectiveness differs based on the choice of first line agent.

email: elizabeth.handorf@fccc.edu

20. FOUNDATIONS OF STATISTICAL INFERENCE IN THE ERA OF MACHINE LEARNING

DEEP FIDUCIAL INFERENCE

Jan Hannig*, University of North Carolina, Chapel Hill
Gang Li, University of North Carolina, Chapel Hill

R. A. Fisher developed the idea of fiducial inference during the first half of the 20th century. While his proposal led to interesting methods for quantifying uncertainty, other statisticians of the time did not fully accept Fisher's approach. Beginning around the year 2000, the idea of fiducial inference was re-investigated and it was discovered that Fisher's approach, when properly generalized, opens doors to solve many important and difficult inference problems. The generalization of Fisher's idea was termed generalized fiducial inference (GFI). The main idea of GFI is to carefully transfer randomness from the data to the parameter space using an inverse of a data generating equation without the use of Bayes theorem. After more than a decade of investigations, a unifying theory for GFI was developed, and GFI solutions to many challenging practical problems in different fields of science and industry were provided. In this talk we discuss interaction of GFI and machine learning, including how certain computations within generalized fiducial framework can be made using an autoencoder.

email: jan.hannig@unc.edu

UNCERTAINTY QUANTIFICATION OF TREATMENT REGIME IN PRECISION MEDICINE BY CONFIDENCE DISTRIBUTIONS

Minge Xie*, Rutgers University
Yilei Zhan, Rutgers University
Sijian Wang, Rutgers University

Personalized decision rule in precision medicine is a discrete parameter, for which theoretical development of statistical inference is lacking. This talk proposes a new way to quantify the estimation uncertainty in a personalized decision based on confidence distribution (CD). Suppose, in a regression setup, the optimal decision for treatment versus control for an individual z is determined by a linear decision rule D = I(m1(z)> m0(z)), where m1(z) and m0(z) are the expectations of potential outcomes of treatment and control, respectively. The estimated D has uncertainty. We propose to find a CD for v = m1(z) - m0(z) and compute a confidence measure of the decision {D=1} = {v>0}. This measure, with value in [0,1], provides a frequency-based assessment about the decision. For example, if the measure for {D=1} is 63%, then, out of 100 patients the same as patient z, 63 will benefit using treatment and 37 will be better off in control group. This confidence measure is shown to match well with the classical assessments of sensitivity and specificity, but without the need to know the true D. Utility of the development is demonstrated in an adaptive clinical trial.

email: mxie@stat.rutgers.edu

HEALTHIER FAST FOOD: AUTOMATED DEBIASING OF BAYESIAN POSTERIORS

Keli Liu*, Stanford University
Xiao-Li Meng, Harvard University

Data analysts today seek algorithms that can handle large amounts of data that often do not fit onto a single machine. And they want these algorithms to run with as little manual input as possible. The goal is to deliver such convenience while ensuring that the inferences are approximately correct—the holy grail of "healthy fast food". Such a paradigm poses a particular challenge for Bayesian inference: (i) limited manual input means one cannot tailor priors to the problem at hand (some sort of default is used, without careful understanding of its implications for the posterior) and (ii) current algorithms for distributed Bayesian computation often result in biased posterior sampling. Motivated by the fiducial philosophy of "continuing to trust", which uses a pre-data pivotal distribution posthoc, we provide a simple and quick way to diagnose the seriousness of these complications. This leads to a method for automatically debiasing the posterior (to first order), with results that are healthier yet still fast.

email: keliliu@stanford.edu

21. CLINICAL TRIALS: CANCER APPLICATIONS AND SURVIVAL ANALYSIS

COMPARISON OF POPULATION REGISTRY OBSERVATIONAL STUDIES AND RANDOMIZED CLINICAL TRIALS IN ONCOLOGY

Holly E. Hartman*, University of Michigan
Payal D. Soni, University of Michigan
Robert T. Dess, University of Michigan
Ahmed Abugharib, University of Michigan
Steven G. Allen, University of Michigan
Felix Y. Feng, University of California, San Francisco
Anthony L. Zietman, Massachusetts General Hospital
Reshma Jagsi, University of Michigan
Daniel E. Spratt, University of Michigan
Matthew J. Schipper, University of Michigan

The number of studies utilizing population-based registries (e.g. SEER) to test treatment effectiveness is increasing substantially. In the Oncology setting, we examine 755 registry studies that compare two or more treatments in terms of overall survival (OS). Reporting quality was poor with only 38% of the observational studies (OBS) reporting median follow up and 63% reporting how missing data was handled. 355 of these studies were matched to 121 randomized control trials (RCTs) comparing the same treatments in the same disease site and using OS as the endpoint. Agreement between the OBS and RCT was defined qualitatively as both studies having the same statistical conclusions. Only 40% of OBS agreed with their matched RCT. Study factors such as data source, year published, and statistical methods used were not associated with agreement. There was no correlation between the hazard ratio reported by the OBS and RCT (Concordance Correlation Coefficient = 0.083, 95% CI = (-0.068, 0.230)). The lack of reporting quality and lack of agreement with RCTs suggest that results from OBS should be utilized with caution, but also motivate rigorous statistical methods to address confounding.

email: holhart@umich.edu

STATISTICAL CONSIDERATIONS FOR TRIALS THAT STUDY SUBPOPULATION HETEROGENEITY

Alexander M. Kaizer*, University of Colorado Denver
Brian P. Hobbs, Cleveland Clinic
Nan Chen, University of Texas MD Anderson Cancer Center
Joseph S. Koopmeiners, University of Minnesota

Breakthroughs in cancer biology have defined new research programs emphasizing the development of therapies that target specific pathways in tumor cells or promote anti-cancer immunity. Innovations in clinical trials have followed with master protocols with inclusive eligibility and designs devised to compare multiple therapies or histologies. However, identifying optimal designs that take into account subpopulation heterogeneity has been challenging, with designs either treating subgroups independently, resulting in reduced power, or pooling all data, ignoring subgroup effects. We propose a new, general framework to identify optimal designs, where subgroup heterogeneity is central to the hypothesis, which accommodates any particular trade-off of operating characteristics while considering multiple scenarios for heterogeneity. The framework is applied to identify optimal basket trial designs that maximize power and monitor the potential for heterogeneity among patients with differing clinical indications among designs that treat baskets independently, pool the data, or use Bayesian multisource exchangeability modeling to enable the sharing of information across baskets.

email: alex.kaizer@ucdenver.edu

DESIGNING CLINICAL TRIALS WITH RESTRICTED MEAN SURVIVAL TIME ENDPOINT AS PRACTICAL CONSIDERATIONS

Anne Eaton*, University of Minnesota
Terry Therneau, Mayo Clinic
Jennifer Le-Rademacher, Mayo Clinic

Background: The difference in restricted mean survival time is gaining popularity as a measure of treatment benefit because it directly measures the quantity of most interest to patients, and retains interpretability under non-proportional hazards. Our goal is to provide researchers with practical guidance to design trials with mean survival time as the primary endpoint. Methods: We compare the power of the mean survival time test and log-rank tests in four scenarios using plots of power versus each design parameter, and simulations. Results: Under proportional hazards, the power of the restricted mean survival time test approaches that of the log-rank test if the restriction time is late. Under non-proportional hazards, it may out-power the log-rank test for some restriction times. Many factors interact to determine the power; we recommend plotting relationships between each design parameter and power in your scenario of interest. We provide software that estimates sample size and generates these plots. Conclusion: Researchers can use the information and tools provided to design trials with restricted mean survival time as the primary endpoint.

email: eato0055@umn.edu

A GENERALIZED PERMUTATION PROCEDURE TO ASSOCIATE PATHWAYS WITH CLINICAL OUTCOMES

Stanley B. Pounds*, St. Jude Children's Research Hospital
Xueyuan Cao, University of Tennessee Health Science Center

In modern molecular clinical oncology research, a common scientific question is to evaluate the association of a pathway or other gene set with clinical outcomes. Several methods exist to evaluate the association of a binary or qualitative outcome with a pathway. However, there are few methods to evaluate the association of a survival (censored time-to-event) endpoint with a gene set. Here, we generalize the multi-response permutation procedure (MRPP; PubMed ID 18042553) to evaluate associations of a gene-set with a quantitative endpoint or a survival endpoint by considering all possible dichotomizations of a quantitative endpoint and all risk-sets for the survival endpoint. Across a series of simulation studies, the generalized MRPP (gMRPP) maintains type I error control. Also, in most settings evaluated by simulation, the gMRPP shows statistical power comparable to or better than other methods to identify associations of a gene-set with a qualitative, quantitative, or survival endpoint. Finally, we show results of the procedure in a pediatric leukemia application.

email: stanley.pounds@stjude.org

BAYESIAN CLINICAL TRIAL DESIGN FOR JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

Jiawei Xu*, University of North Carolina, Chapel Hill
Matthew A. Psioda, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Joint models for longitudinal and survival data have become mainstream tools for analyzing time-to-event data in clinical trials, but very few methods exist for designing clinical trials using these models. We develop a Bayesian design methodology based on Bayesian versions of type I error and power that are defined with respect to the posterior distribution for the parameters and conditional on the relevant hypothesis being true. We demonstrate that when the design controls the Bayesian type I error rate, meaningful amounts of information can be borrowed from the prior and that the borrowable amount increases with the sample size in the new trial. In contrast, when the design is required to control type I error in the traditional frequentist sense, all prior information must be discarded. We introduce a connection between Bayesian analyses with power prior and a weighted maximum likelihood analysis. This leads to an accurate approximation for the posterior probabilities required for Bayesian inference and obviates the need for Markov Chain Monte Carlo methods for Bayesian model fitting. We demonstrate our novel methodology by designing a clinical trial in breast cancer.

email: jiawei@live.unc.edu

INTERPRETATION OF TIME-TO-EVENT OUTCOMES IN RANDOMIZED TRIALS

Ludovic Trinquart*, Boston University School of Public Health
Isabelle Weir, Boston University School of Public Health

Background: Multiple features in the presentation of randomized controlled trial (RCT) results are known to influence comprehension and interpretation. Methods: We performed a randomized experiment. We selected 15 cancer RCTs. We created 3 versions reporting either the hazard ratio (HR), difference in restricted mean survival times (RMSTD), or HR+RMSTD for the primary outcome. We randomized participants to one of 15 abstracts and one of 3 versions. We asked how beneficial the experimental treatment was (0 to 10 Likert scale). We also asked a multiple-choice question about interpretation of the HR. Results: We randomly allocated 469 participants. The mean Likert score was statistically significantly lower in the RMSTD group vs. the HR group (mean difference -0.8, 95% confidence interval, -1.3 to -0.4, $p<0.01$) and vs. the HR+RMSTD group (-0.6, -1.1 to -0.1, $p=0.05$). In all, 47.2% (42.7% to 51.8%) of participants misinterpreted the HR, with 40% equating it with a reduction in absolute risk. Conclusion: Misinterpretation of the HR is common. Participants judged experimental treatments to be less beneficial when presented with RMSTD as compared with HR.

email: ludovic@bu.edu

22. MULTIPLE TESTING

A BOTTOM-UP APPROACH TO TESTING HYPOTHESES THAT HAVE A BRANCHING TREE DEPENDENCE STRUCTURE, WITH FALSE DISCOVERY RATE CONTROL

Yunxiao Li*, Emory University
Yijuan Hu, Emory University
Glen A. Satten, Centers for Disease Control and Prevention

Modern statistical analyses often involve testing large numbers of hypotheses. In many situations, these hypotheses may have an underlying tree structure that not only helps determine the order that tests should be conducted but also imposes a dependency

between tests that must be accounted for. Our motivating example comes from testing the associations between groups of microbes (organized in operational taxonomic units or OTUs) and a trait of interest. Given p-values from association tests for each individual OTU, we would like to know if we can declare that a certain species, genus, or higher taxonomic grouping can be considered to be associated with the trait. We develop a bottom-up testing algorithm that controls the error rate of decisions made at higher levels in the tree, conditioning on findings at lower levels in the tree. We further show this algorithm controls the false discovery rate based on the global null hypothesis that no taxa are associated with the trait. By simulation, we also show that our approach is better at finding driver taxa, the highest level taxa for which there are dense association signals in all taxa below the driver taxon.

email: yunxiao.li@emory.edu

A GLOBAL HYPOTHESIS TEST FOR DEPENDENT ENDPOINTS BASED ON RESEARCHER'S PREDICTIONS

Robert N. Montgomery*, University of Kansas Medical Center
Jonathan Mahnken, University of Kansas Medical Center

The issue of multiplicity has attracted a great deal of study, with advances resulting in increases in power and Type I error control for complicated sets of endpoints. However there are many situations in biomedical research where a hypothesis is of interest but cannot be directly evaluated, thus many endpoints that partially address the hypothesis are used as proxies. The goal in such experiments is often to reject or fail to reject a global hypothesis based on the results of the endpoints, this is often by using primary and secondary endpoints. In these situations if one of the primary endpoints is statistically significant we would reject the overall hypothesis. Unfortunately this methods and many others often cannot directly answer the research question of interest. We derive a new hypothesis test that addresses a global hypothesis based on many dependent endpoints. Our test is based on the correlation matrix between endpoints and a vector of researcher's predictions. We show that our method has good power and Type I error control.

email: robertnmontgomery@gmail.com

PER-FAMILY ERROR RATE CONTROL FOR GAUSSIAN GRAPHICAL MODEL VIA KNOCKOFFS

Siliang Gong*, University of Pennsylvania
Qi Long, University of Pennsylvania
Weijie Su, University of Pennsylvania

Driven by many real applications including, but not limited to, estimation of biological pathways, the estimation of and inference for Gaussian Graphical Models (GGM) are fundamentally important and have attracted substantial research interest in the literature. However, it is still challenging to achieve overall error rate control when recovering the graph structures of GGM. In this paper, we propose a new multiple testing method for GGM using the knockoffs framework introduced by Barber and Candès. Our method is shown to control the overall finite-sample Per-Family Error Rate up to a probability error bound induced by the estimation errors of knockoff features. The performance of our method is evaluated in extensive numerical studies.

email: siliang@pennmedicine.upenn.edu

PERMUTATIONS UNLOCK EXPERIMENT-WISE NULL DISTRIBUTIONS FOR LARGE SCALE SIMULTANEOUS INFERENCE OF MICROBIOME DATA

Stijn Hawinkel*, Ghent University, Belgium
Luc Bijnens, Janssen Pharmaceutical companies of Johnson and Johnson, Belgium
Olivier Thas, Ghent University, Belgium

Correlation between statistical tests complicates multiplicity correction, as it inflates the variability of the false discovery proportion (FDP), i.e. the fraction of false discoveries among the rejected null hypotheses. This is because correlation causes the observed distribution of the test statistics to depart from the theoretical null distribution. Moreover, the null distribution of the ensemble of test statistics is unique for every experiment and does not necessarily follow a known distribution. We argue that it is crucial for accurate multiplicity correction to estimate this experiment-wise null distribution. We achieve this by generating many permutation distributions, and estimating the experiment-wise null distribution as a weighted average, with weights based on similarity with the observed distribution of test statistics. Unlike existing approaches, our method does not require knowledge on the correlation structure or on the shapes of the null distributions. Simulations show that our method reduces the variability of the FDP, while controlling the FDR and increasing the sensitivity.

email: stijn.hawinkel@ugent.be

GLOBAL AND SIMULTANEOUS HYPOTHESIS TESTING FOR HIGH-DIMENSIONAL LOGISTIC REGRESSION MODELS

Rong Ma*, University of Pennsylvania
Tony T. Cai, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

High-dimensional logistic regression is widely used in analyzing data with binary outcomes. In this paper, global testing and large-scale multiple testing for the regression coefficients are considered. A test statistic for testing the global null hypothesis is constructed using a generalized low-dimensional projection method for bias correction and its asymptotic null distribution is derived. For testing the individual coefficients simultaneously, a multiple testing procedure is proposed and shown to control the false discovery rate (FDR) asymptotically. Simulation studies are carried out to examine the numerical performance of the proposed tests and their superiority over existing methods. The testing procedures are also illustrated by analyzing a metabolomics study that investigates the association between fecal metabolites and pediatric Crohn's disease and the effects of treatment on such associations.

email: rongm@upenn.edu

IDENTIFYING RELEVANT COVARIATES IN RNA-Seq ANALYSIS BY PSEUDO-VARIABLE AUGMENTATION

Yet Nguyen*, Old Dominion University
Dan Nettleton, Iowa State University

RNA-sequencing (RNA-seq) technology enables the detection of differentially expressed genes, i.e., genes whose mean transcript abundance levels vary across conditions. In practice, an RNA-seq dataset often contains some explanatory variables that will be included in analysis with certainty in addition to a set of

covariates that are subject to selection. Some of the covariates may be relevant to gene expression levels, while others may be irrelevant. Either ignoring relevant covariates or attempting to adjust for the effect of irrelevant covariates can be detrimental to identifying differentially expressed genes. We address this issue by proposing a covariate selection method using pseudo-covariates to control the expected proportion of selected covariates that are irrelevant. We show that the proposed method can accurately choose the most relevant covariates while holding the false selection rate below a specified level. We also show that our method performs better than methods for detecting differentially expressed genes that do not take covariate selection into account, or methods that use surrogate variables instead of the available covariates.

email: ynguyen@odu.edu

ASYMPTOTIC SIMULTANEOUS CONFIDENCE INTERVALS OF ODDS RATIO IN MANY-TO-ONE COMPARISON OF PROPORTIONS FOR CORRELATED PAIRED BINARY DATA

Xuan Peng*, State University of New York at Buffalo
Chang-Xing Ma, State University of New York at Buffalo

In many medical researches, measurements obtained from paired organs (e.g., eyes or ears) of an unit are generally highly correlated. It is very important to account for the intraclass correlation on statistical inferences, since ignoring the intraclass correlation between paired measurements may yield biased inferences. In addition, it is commonly needed to consider simultaneous comparison of proportions of success between a single control group and multiple treatment groups in randomized clinical trials. In this research, we constructed simultaneous confidence intervals (SCIs) for odds ratio in a many-to-one comparison framework under such correlated paired binary data. Four different methods are applied to construct simultaneous confidence interval for odds ratio with Dunnett-like or Bonferroni multiple adjustment. The empirical coverage probabilities and mean interval widths of the SCIs from resulting methods are compared through a Monte Carlo simulation study to evaluate their performance. A real work example is included to illustrate the usage of the resulting methods.

email: xuanpeng@buffalo.edu

23. CLUSTERED DATA METHODS

SAMPLE SIZE ESTIMATION FOR STRATIFIED INDIVIDUAL AND CLUSTER RANDOMIZED TRIALS WITH BINARY OUTCOMES

Lee Kennedy-Shaffer*, Harvard University
Michael D. Hughes, Harvard T.H. Chan School of Public Health

Individual randomized trials (IRTs) and cluster randomized trials (CRTs) with binary outcomes arise in a variety of settings and are often analyzed by logistic regression and generalized estimating equations with a logit link, respectively, but the effect of stratification on the required sample size is not well understood for these methods. We propose easy-to-use methods for sample size estimation for stratified IRTs and CRTs and identify the ratio of the sample size for a stratified trial versus a comparably-powered unstratified trial, allowing investigators to evaluate how stratification will affect the sample size when planning a trial. Using these methods, we describe scenarios where stratification may have an important impact on the required sample size. In the two-stratum case, there are unlikely to be plausible scenarios in which an important sample size reduction is achieved when the overall probability of the outcome is low, both for IRTs and for CRTs with small cluster sizes. When the probability of events is not small, or when cluster sizes are large, however, there are scenarios where practically important reductions in sample size result from stratification.

email: lee_kennedyshaffer@g.harvard.edu

SAMPLE SIZE CONSIDERATIONS FOR STRATIFIED CLUSTER RANDOMIZATION DESIGN WITH BINARY OUTCOMES AND VARYING CLUSTER SIZE

Xiaohan Xu*, University of Texas Southwestern Medical Center and Southern Methodist University
Hong Zhu, University of Texas Southwestern Medical Center
Chul Ahn, University of Texas Southwestern Medical Center

Stratified cluster randomization trials (CRTs) are frequently employed in clinical and healthcare research. Comparing to simple randomized CRTs, stratified CRTs reduce the imbalance of baseline prognostic factors among different intervention groups. Despite the popularity, there is limited methodological development on sample size estimation for stratified CRTs, and existing work mostly assumes equal cluster size within each stratum. Clusters are often naturally formed with random sizes in CRTs. With varying cluster size, commonly used approaches ignore the variability in cluster size, which may underestimate (overestimate) the required sample size and lead to underpowered (overpowered) clinical trials. We propose a closed-form sample size formula for stratified CRTs with binary outcomes, accounting for both clustering and varying cluster size. We investigate the impact of various design parameters on the relative change in sample size due to varying cluster size. Simulation studies are conducted and an application to a pragmatic trial of a triad of chronic kidney disease, diabetes and hypertension is presented for illustration.

email: Xiaohan.Xu@UTSouthwestern.edu

POWER CALCULATION FOR STEPPED WEDGE DESIGNS WITH BINARY OUTCOMES

Xin Zhou*, Harvard T. H. Chan School of Public Health
Xiaomei Liao, AbbVie Inc.
Donna Spiegelman, Yale School of Public Health

The stepped wedge design (SWD) is a relatively new type of cluster trial design, and it is increasing in popularity in public health and implementation science. In an SWD, every cluster begins in the control condition; clusters are then randomly selected at specified time steps to switch from the control to the intervention. Every cluster receives the intervention by the end of the study. Although most outcomes in health care trials are binary, power calculations for SWDs of binary outcomes have not been well investigated. A maximum likelihood method was recently developed for binary outcomes with an identity link. However, the logit link, which is the canonical link in logistic regression, is dominant in relevant applications. In this work, we extend the maximum likelihood method by applying the logit link for power calculations. We studied the robustness of the power to the varying cluster size and to the distribution of the cluster random effects, and compare

power for the two link functions. We applied the new method to the design of a large-scale intervention program on postpartum intra-uterine device insertion services in Tanzania.

email: stxzh@channing.harvard.edu

PAN-DISEASE CLUSTERING ANALYSIS OF THE TREND OF PERIOD PREVALENCE

Chenjin Ma*, Renmin University of China
Sneha Jadhav, Yale University
Ben-Chang Shia, Taipei Medical University
Shuangge Ma, Yale University

For all diseases, prevalence has been carefully studied. In the "classic" paradigm, the prevalence of different diseases has usually been studied separately. Accumulating evidences have shown that diseases can be "correlated". The joint analysis of prevalence of multiple diseases can provide important insights beyond individual-disease analysis, however, has not been well conducted. In this study, we take advantage of the uniquely valuable Taiwan National Health Insurance Research Database. The goal is to identify clusters within which diseases share similar period prevalence trends. For this purpose, a novel penalization pursuit approach is developed, which has an intuitive formulation and satisfactory properties. In data analysis, the period prevalence values are computed using records on close to 1 million subjects and 14 years of observation. For 405 diseases, 35 nontrivial clusters and 27 trivial clusters are identified. A closer examination suggests that the clustering results have sound interpretations. This study is the first to conduct a pan-disease clustering analysis of prevalence trend using the uniquely valuable NHIRD and can have important value in multiple aspects.

email: chenjin.ma@yale.edu

ORDINAL CLUSTERED DATA WITH INFORMATIVE CLUSTER SIZE IN A LONGITUDINAL STUDY

Aya A. Mitani*, Boston University
Elizabeth K. Kaye, Boston University
Kerrie P. Nelson, Boston University

Previous work on modeling marginal inference in longitudinal studies with informative cluster size (ICS) has focused on continuous outcomes. However, in a longitudinal study of periodontal disease, the outcome was measured using an ordinal scoring system. In addition, participants may lose teeth over the course of the study due to advancing disease status. Here we develop longitudinal cluster weighted generalized estimating equations (CWGEE) to model the association of ordinal clustered longitudinal outcomes with participant-level health-related covariates including metabolic syndrome and smoking status and potentially decreasing cluster size due to loss of teeth over time, by fitting a proportional odds logistic regression model. The within-teeth correlation coefficient over time is estimated using the two-stage quasi-least squares method. In an extensive simulation study, we compare results obtained from CWGEE with various working correlation structures to those obtained from unweighted GEE which does not account for ICS. Our proposed method yields results with very low biases and excellent coverage probabilities in contrast to unweighted GEE.

email: amitani@bu.edu

INFERENTIAL PROCEDURES FOR CONSENSUS CLUSTERING- AN ANOVA BASED APPROACH

Kenneth T. Locke, Jr.*, University of Pennsylvania
Yong Chen, University of Pennsylvania
J. Richard Landis, University of Pennsylvania

With rising interest in precision medicine, cluster discovery methods are crucial in forming subgroups of individuals. Consensus Clustering utilizes multiple runs of a clustering algorithm and generates a matrix of pairwise comparisons to provide a proportion of how often a pair of individuals are found in the same cluster. Through methods such as the Proportion of Ambiguously Clustered pairs (PAC) measure developed by Senbabaoglu et al. in 2014, we can roughly determine the ideal number of clusters K. However, the PAC was not developed under a hypothesis testing framework and therefore lacks a p-value or test statistic for deciding how many clusters exist in the data. In our method, we will be using the pairwise consensus matrix of patients to develop a hypothesis test with the null being that no clusters are present and the alternative that K clusters are present. Utilizing permutation tests and Edgeworth expansion, we estimate the p-value under the null hypothesis and find the most significant K indicating the ideal number of clusters. From simulated and real data, we find that our method is accurate in predicting the correct number of clusters.

email: lockek@pennmedicine.upenn.edu

INFORMATIVELY EMPTY CLUSTERS WITH APPLICATION TO TRANSGENERATIONAL STUDIES

Glen W. McGee*, Harvard University
Marc G. Weisskopf, Harvard University
Marianthi-Anna Kioumourtzoglou, Columbia University
Brent A. Coull, Harvard University
Sebastien Haneuse, Harvard University

Exposures that act epigenetically can affect increasingly more people as the exposed population reproduces. Transgenerational studies, however, are susceptible to informative cluster size, occurring when the number of children to a mother (the cluster size) is related to their outcomes, given covariates. But what if some women bear no children at all? The impact of these potentially informative empty clusters is yet unknown. We evaluate the performance of standard methods for informative cluster size when cluster size is permitted to be zero. We find that if the informative cluster size mechanism induces empty clusters, standard methods lead to biased parameter estimates. Joint models of outcome and size permit valid conditional inference if empty clusters are explicitly included in the analysis, but in practice they regularly go unacknowledged. By contrast, estimating equation approaches omit empty clusters and yield biased estimates of marginal effects, and we propose a joint marginalized approach to incorporate empty clusters. Competing methods are compared via simulation and in a study of diethylstilbestrol exposure and ADHD among 106,198 children to 47,540 nurses.

email: glenmcgee@g.harvard.edu

24. GENOME WIDE ASSOCIATION STUDIES AND OTHER GENETIC STUDIES

G-SMUT: GENERALIZED MULTI-SNP MEDIATION INTERSECTION-UNION TEST

Wujuan Zhong*, University of North Carolina, Chapel Hill
Cassandra N. Spracklen, University of North Carolina, Chapel Hill
Karen L. Mohlke, University of North Carolina, Chapel Hill
Xiaojing Zheng, University of North Carolina, Chapel Hill
Jason Fine, University of North Carolina, Chapel Hill
Yun Li, University of North Carolina, Chapel Hill

To elucidate mechanisms behind GWAS (Genome-Wide Association Studies) SNPs (Single Nucleotide Polymorphisms) identified for a variety of phenotypic outcome such as disease status and microbiome count measurements from sequencing data, we propose to use IUT (Intersection-Union Test) combined with LRT (Likelihood Ratio Test) to detect mediation effect of multiple SNPs via some mediator (for example, the expression of a neighboring gene) on outcome, where IUT is mainly used to decompose mediation effect and LRT to test coefficient of mediator conditional on SNPs. Under the alternative (or null) hypothesis, we fit high-dimensional generalized linear mixed models under mediation framework. Laplace approximation is applied to compute the marginal likelihood of outcome and coordinate descent algorithm is used to estimate corresponding parameters. Our extensive simulations demonstrate the validity of our proposed method and substantial, up to 97%, power gains over alternative methods. We believe our proposed method will be a useful tool in this post-GWAS era to disentangle the potential causal mechanism from DNA to phenotype for a new drug discovery and personalized medicine.

email: zhongwujuan@gmail.com

A MULTIPLE-WEIGHTED FALSE DISCOVERY RATE CONTROLLING PROCEDURE IN GENOME-WIDE ASSOCIATION STUDIES

Zhou Fang*, Brigham and Women's Hospital
Nikolaos Patsopoulos, Brigham and Women's Hospital and Broad Institute

In the past decade, genome-wide association studies (GWAS) have allowed the unprecedented investigation of the genome's contribution to various traits, identifying more than 60,000 robust associations that span over 2,400 traits or diseases. With increased resolution, current GWAS involve the analysis of millions of genetic variants of variable minor allele frequency (MAF) that are at some degree correlated, i.e. are in linkage disequilibrium (LD), posing serious challenges in controlling for the false discovery rate (FDR). However, neither current FDR controlling methods nor fixed type I error rate threshold, i.e. 5x10-8, properly accounts either for the variance of MAF or LD in a genome, when both may lead to differences in power to detect a true effect. To address these issues, we propose an FDR controlling procedure weighted on both LD and MAF. We compare our method with typically used multiplicity adjusting approaches in simulated and actual GWAS data. The proposed method provided better balance in discovering variants with different MAF pattern, and attenuated the loss of power due to the nature that multiple tests are highly correlated.

email: zfang4@bwh.harvard.edu

PHENOME-WIDE SNP-SET ASSOCIATION TEST BASED ON GWAS SUMMARY DATA TO IDENTIFY NOVEL DISEASE-GENE ASSOCIATION

Bin Guo*, University of Minnesota
Baolin Wu, University of Minnesota

We study statistical methods for SNP-set association test with multiple traits using only GWAS summary data. The proposed methods PheSATS (Phenome-wide SNP-set Association Test using Summary data) further generalize the phenome-wide association study (PheWAS) by integrating both multiple traits and multiple variants to detect novel disease-gene association. PheSATS capitalizes on the polygenic nature of most human traits and the pleiotropy association of variants to detect novel genetic variants. We provide transparent derivations to rigorously show the statistical justifications for the PheSATS methods. We develop efficient numerical algorithms to accurately compute analytical p-values without computing-intensive resampling. PheWAS are applicable to multiple GWAS measured on one cohort or from studies with arbitrarily overlapped samples including independent GWAS. PheSATS have well-controlled type I errors verified through the simulation study. We further apply PheSATS to multiple lipids and psychiatric disorders GWAS summary data. We found many novel loci that were not detected by existing single-trait single-SNP based tests, and are worth further study.

email: guoxx617@umn.edu

POLYGENIC RISK PREDICTION USING FUNCTIONAL ANNOTATION: APPLICATION TO THE INTERNATIONAL LUNG CANCER CONSORTIUM (ILCCO)

Jingwen Zhang*, Harvard T.H. Chan School of Public Health
Xihong Lin, Harvard T.H. Chan School of Public Health

Lung cancer is a well-known leading cause of death worldwide. Previous studies have revealed the polygenic structure of such diseases, which motivate us to quantify lifetime risk using genome-wide data. To assess the genetic risk in general population, we construct a polygenic prediction model using the GWAS summary statistics from the International Lung Cancer Consortium (ILCCO). To further improve the prediction power, we evaluated the enrichment of functionally annotated SNPs in GWAS and included such information in the prediction model. Finally, we compared our method with existing models and discussed the necessity of including functional annotation in lung cancer risk prediction.

email: jingwn.zhang@gmail.com

A SIMPLE AND GENERAL COLOCALIZATION TEST

Yangqing Deng*, University of Minnesota
Wei Pan, University of Minnesota

Testing colocalization of GWAS causal variants and eQTL causal variants can help establish causal relationships: if a GWAS trait and a gene's expression share the same causal variant, then it may suggest a regulatory role of the causal SNP on gene expression to the trait. Accordingly, it is of interest to develop various colocalization testing approaches. The existing approaches all have some severe limitations. Some use the null hypothesis that there is colocalization, which may limit the statistical power. Some restrict the number of causal SNPs in a locus, which may lead to loss

of power in the presence of allelic heterogeneity. Most methods cannot be applied to summary statistics or cases with more than two traits. We develop a simple and general approach based on conditional modeling, which overcomes the above problems. We demonstrate that compared with other methods, our new method can be applied to a wider range of cases and may perform better in certain scenarios, using both simulated and real data.

email: yangq001@umn.edu

INTEGRATIVE GENE-BASED ASSOCIATION TESTING FOR CANCER PHENOTYPES WITH SOMATIC TUMOR EXPRESSION DATA AND GWAS SUMMARY DATA

Jack W. Pattee*, University of Minnesota
Wei Pan, University of Minnesota

Genome-wide association studies (GWAS) have successfully identified many genetic variants associated with complex traits. However, GWAS experience power issues, resulting in the failure to detect certain associated variants. Additionally, GWAS are often unable to parse the biological mechanisms driving associations. Existing gene-based association test TWAS leverages eQTL data to increase the power of association tests and illuminate the biological mechanisms by which genetic variants modulate complex traits. We extend the TWAS methodology to cancer phenotypes. Our methodology combines somatic data from tumor cells with germline genetic information from matched normal tissue cells, allowing us to leverage information from the nuanced somatic landscape of tumor cells. We use somatic and germline data on lung adenocarcinomas from The Cancer Genome Atlas in conjunction with a meta-analyzed lung cancer GWAS to identify novel genes associated with lung cancer.

email: patte631@umn.edu

25. TIME SERIES

EMPIRICAL LOCALIZED TIME-FREQUENCY ANALYSIS VIA PENALIZED REDUCED RANK REGRESSION

Marie Tuft*, University of Pittsburgh
Robert Todd Krafty, University of Pittsburgh

Spectral analysis of nonstationary biological processes such as heart rate variability (HRV) and EEG poses a unique challenge: localization and accurate descriptions of both frequency and time are required. By reframing this question in a reduced rank regression setting, we propose a novel approach that produces a low-dimensional and interpretable empirical basis localized in time and frequency. To estimate this frequency-time basis, first we partition the time series into n stationary intervals and calculate the periodogram at each of the m Fourier frequencies. Then we use reduced rank regression with singular value decomposition on the resulting n x m matrix. An adaptive sparse fused lasso penalty is applied to the estimates of the left and right singular vectors which ensures localization and smoothness in both frequency and time. Asymptotic properties of this method are derived, and it is shown to provide a consistent estimator of the time-varying spectrum. Simulation studies are used to evaluate its performance and its utility in practice in illustrated through the analysis of HRV during sleep.

email: marie.tuft@pitt.edu

MULTI-SUBJECT SPECTRAL ANALYSIS OF RESTING-STATE EEG SIGNALS FROM TWINS USING A NESTED BERNSTEIN DIRICHLET PRIOR

Brian B. Hart*, University of Minnesota
Michele Guindani, University of California, Irvine
Stephen Malone, University of Minnesota
Mark Fiecas, University of Minnesota

Electroencephalography (EEG) is a non-invasive neuroimaging modality that captures electrical brain activity many times per second. We seek to estimate power spectra from EEG data that was gathered for 557 adolescent twin pairs through the Minnesota Twin Family Study (MTFS). Typically, spectral analysis methods treat time series from each subject separately, and independent spectral densities are fit to each time series. We take advantage of the twin design of the study by borrowing information across subjects in order to model and conduct statistical inference on the spectral densities of the EEG signals. To this end, we propose a Nested Bernstein Dirichlet Prior model to estimate the power spectrum of the EEG signal for each subject. Our method estimates spectral densities through data driven smoothing of periodograms within and across subjects while requiring minimal user input to tuning parameters. The method also facilitates heritability analyses on features of the estimated spectral density curves such as peak frequency and frequency band power.

email: bbhart06@gmail.com

EMPIRICAL FREQUENCY BAND ANALYSIS OF NONSTATIONARY TIME SERIES

Scott A. Bruce*, George Mason University
Cheng Yong Tang, Temple University
Martica H. Hall, University of Pittsburgh
Robert T. Krafty, University of Pittsburgh

The time-varying power spectrum of a time series process quantifies the magnitude of oscillations at different frequencies and times. To obtain low-dimensional, parsimonious measures from this functional parameter, applied researchers consider collapsed measures of power within local bands of frequencies. Frequency bands commonly used in the scientific literature were historically derived, but they are not guaranteed to be optimal or justified for adequately summarizing information from a given time series. There is a dearth of methods for empirically constructing statistically optimal bands for a given signal. We seek to provide a standardized, unifying approach for deriving and analyzing customized frequency bands. A consistent, frequency-domain, iterative cumulative sum based scanning procedure is formulated to identify frequency bands that best preserve nonstationary information. A formal testing procedure is also developed to test which, if any, frequency bands remain stationary. The proposed method is used to analyze heart rate variability of a patient during sleep and uncovers a refined partition of frequency bands that best summarize the time-varying power spectrum.

email: sbruce7@gmu.edu

JOINT STRUCTURAL BREAK DETECTION AND PARAMETER ESTIMATION IN HIGH-DIMENSIONAL NON-STATIONARY VAR MODELS

Abolfazl Safikhani*, Columbia University
Ali Shojaie, University of Washington, Seattle

Assuming stationarity is unrealistic in many time series applications. A more realistic alternative is to assume piecewise stationarity, where the model is allowed to change at potentially many time points. We propose a three-stage procedure for consistent estimation of both structural change points and parameters of high-dimensional piecewise vector autoregressive (VAR) models. In the first step, we reformulate the change point detection problem as a high-dimensional variable selection one, and solve it using a penalized least square estimator with a total variation penalty. We show that the proposed penalized estimation method over-estimates the number of change points. We then propose a selection criterion to identify the change points. In the last step of our procedure, we estimate the VAR parameters in each of the segments. We prove that the proposed procedure consistently detects the number of change points and their locations. We also show that the procedure consistently estimates the VAR parameters. The performance of the method is illustrated through several simulation studies and real data examples.

email: as5012@columbia.edu

ORDER RESTRICTED INFERENCE IN CHRONOBIOLOGY

Yolanda Larriba*, University of Valladolid, Spain
Cristina Rueda, University of Valladolid, Spain
Miguel A. Fernández, University of Valladolid, Spain
Shyamal D. Peddada, University of Pittsburgh

This paper is motivated by applications in oscillatory systems where researchers are interested in discovering components displaying rhythmic temporal patterns. The contributions are twofold. First, methodology is developed based on a \textit{circular signal} plus error model defined using order restrictions. This formulation of rhythmicity is easily interpretable and flexible. Second, using Order Restricted Inference based methods, we address problems in oscillatory systems data analysis. Specifically, we develop methodology for detecting rhythmic signals, especially when sample times are unknown. The methodology is computationally efficient, outperforms the existing ones and is applicable to address a wide range of questions in oscillatory systems.

email: yolandalago@hotmail.com

DYNAMIC BAYESIAN PREDICTION AND CALIBRATION USING MULTIVARIATE SENSOR DATA STREAMS

Zhenke Wu*, University of Michigan
Timothy NeCamp, University of Michigan
Srijan Sen, University of Michigan

There is a critical need to understand the temporal dynamics of depression using real-time, objective measures. We introduce a flexible multivariate time series model to analyze multiple sensor data streams collected at distinct time scales (minute, daily, and quarterly) with occasional missingness (due to failure to wear wristbands or carry smartphones). Our model predicts interns' mood and estimates the profile of lagged effects for each predictor time series by sharing information both across time, to account for smooth time-varying associations, as well as across similar subjects. We illustrate our methods using data from the 2017-18 Intern Health Study cohort recruited at University of Michigan. Lastly, we discuss computational issues and the practical implications of our results in the analysis of emerging intensive longitudinal data in mobile health.

email: zhenkewu@umich.edu

26. DOUGLAS ALTMAN: A CONSUMMATE MEDICAL STATISTICIAN

WHAT MAKES A GREAT MEDICAL STATISTICIAN? THE MODEL OF DOUG ALTMAN

Steven Goodman*, Stanford University

This presentation will review the contributions to biostatistical methods, communication and teaching of Doug Altman, 1948-2018, one of the most prolific and influential medical statisticians of his time, despite not having earned an academic doctoral degree, nor having a test, distribution or theorem bearing his name (notwithstanding the Bland-Altman plot). We will discuss what current biostatisticians could learn from his career and from the issues he chose to champion.

email: steve.goodman@stanford.edu

REMEMBERING PROFESSOR DOUG ALTMAN

David Moher*, Ottawa Hospital Research Institute

Doug Altman's career is peppered with warnings about the consequences of inadequate reporting of biomedical research. This call to arms to improve the reporting of medical research is epitomized with his 1994 publication "The scandal of poor medical research" published in the BMJ. Doug's career is also full of examples of him supervising and/or mentoring others to conduct research to provide the required evidence about the associations of biased methods and reporting. Much of this ground-breaking research is still highly cited. I met Doug in the early 1990s. Drummond Rennie, the former deputy editor at JAMA, secured our working relationship for the subsequent 24 years. Doug provided many insights and innovations to the development of CONSORT and reporting guidelines more generally. He developed the CONSORT explanation and elaboration document as a pedagogical tool, now used as a template by many other reporting guideline developers. He also established the EQUATOR Network, in 2006, as a broader generalization of the CONSORT initiative. Doug participated in several studies evaluating the benefits of CONSORT and other reporting guidelines.

email: dmoher@ohri.ca

DOUGLAS ALTMAN AND HIS MENTORING LEGACY

Tianjing Li*, Johns Hopkins Bloomberg School of Public Health

The late Professor Douglas Altman touched the lives of many around the world. Beyond his extraordinary contributions to medical statistics, he mentored hundreds through remarkable independent research careers. As many would agree, he was the most down-to-earth research giant who always made time to encourage, support, and inspire anyone and everyone who knocked on his door. In this presentation, we will share stories from his mentees and mentoring lessons learned through Professor Altman's actions and words.

email: tli19@jhu.edu

27. STATISTICAL ADVANCES FOR EMERGING ISSUES IN HUMAN MICROBIOME RESEARCHES

OPTIMAL PERMUTATION RECOVERY AND ESTIMATION OF BACTERIAL GROWTH DYNAMICS

Hongzhe Li*, University of Pennsylvania
Yuan Gao, University of Pennsylvania

Accurately quantifying microbial growth dynamics for species without complete genome sequences is biologically important but computationally challenging in metagenomics. Here we present DEMIC, a new multi-sample algorithm based on contigs and coverage values, to infer relative distances of contigs from replication origin and to accurately estimate and compare bacterial growth rates between samples. We demonstrate robust performances of DEMIC for a wide range of sample sizes and assembly qualities using various synthetic and real data sets. We provide theoretical analysis to explain why DEMIC works in the framework of optimal permutation recovery.

email: hongzhe@pennmedicine.upenn.edu

HIGHER CRITICISM GOODNESS-OF-FIT TESTS IN PHYLOGENETIC TREES FOR MICROBIOME SEQUENCING EXPERIMENTS

Jeffrey C. Miecznikowski*, State University of New York at Buffalo (SUNY)
Jiefei Wang, State University of New York at Buffalo (SUNY)

There has been an explosion of research in the human microbiome due to advancements in 16S sequencing technology. This has led to big undertakings such as the NIH Human Microbiome Project and other large consortium datasets. Analysis of these data usually involve fitting generalized linear models for each organizational taxonomic unit (OTU) at a phylogenetic level and examining the contrasts for significance. Researchers and clinicians naturally want to analyze the OTUs from the finest resolution of the phylogenetic tree in order to yield the most specific conclusions. Here we use a data-driven approach with exact goodness-of-fit tests based on higher criticism statistics to examine the levels of the phylogenetic tree that may contain the most interesting results.

email: jcm38@buffalo.edu

ANALYZING MATCHED SETS OF MICROBIOME DATA

Yi-Juan Hu*, Emory University
Glen A. Satten, Centers for Disease Control and Prevention
Zhengyi Zhu. Emory University

Matched data arise frequently in microbiome studies. For example, we may have gut microbiome data pre- and post-treatment from a set of individuals, or longitudinal microbiome samples (e.g., vaginal microbiome samples collected in each trimester of a pregnancy). We present a Linear Decomposition Model (LDM) that provides both global test of any effect of the microbiome on the trait of interest and tests of the effects of individual OTUs with false discovery rate (FDR)-based correction for multiple testing. The baseline microbiome characterizing a set is treated as a "nuisance parametes", allowing all efforts to focus on the common differences within sets. The sample correlations are accounted for via block-structured permutation. We evaluate size and power of the global test and FDR and sensitivity of the OTU detection for a wide range of simulated matched-set data and compare to existing methods. We also explore power of different studies, such as with matched and unmatched samples, and provide practical guidance to study designs.

email: yijuan.hu@emory.edu

TESTING STATISTICAL INTERACTIONS BETWEEN MICROBIOME COMMUNITY PROFILES AND COVARIATES

Michael C. Wu*, Fred Hutchinson Cancer Research Center

Microbiome profiling studies are being conducted to find associations between bacterial taxa and a wide range of different outcomes. However, the dimensionality, compositionality, inherent biological structure, and limited availability of samples pose significant challenges. Community level analysis, wherein the entire profile is assessed for association with outcomes, can resolve some of these difficulties but does not easily generalize to analyzing effect modification due to bias incurred in estimating main effects. Thus, under the semi-parametric kernel machine testing framework, we propose a new framework for interaction testing at the community level that incorporates bias reduction approaches in estimating main effects while flexibly capturing interaction terms. Simulations and real data analyses show that our approach correctly controls type I error while maintaining power under a range of situations.

email: mcwu@fhcrc.org

28. WEARABLE TECHNOLOGY IN LARGE OBSERVATIONAL STUDIES

POTENTIAL BATCH EFFECTS AND BIASES IN THE UK BIOBANK ACCELEROMETER DATA

John Muschelli*, Johns Hopkins University

Data from wearable technology, especially wrist-worn accelerometers is becoming more available. Over 100,000 people in a subset of the UK Biobank data set wore accelerometers for approximately 7 days in the study. We explore the potential biases

in the data, including differences in the cohort with data and those without. We also discuss potential calibration and data normalization issues that are within the data that are device-dependent. We apply standard batch-effect correction methods, such as Tukey's biweight, to determine how these affect the overall results in a standard mortality analysis.

email: muschellij2@gmail.com

FUNCTIONAL AND COMPOSITIONAL APPROACHES FOR ACCELEROMETRY WITH APPLICATION TO THE WOMEN'S HEALTH INITIATIVE

Chongzhi Di*, Fred Hutchinson Cancer Research Center

Accelerometers are often used to objectively measure physical activity in epidemiological studies. In these studies, physical activity is often classified into three intensity categories and it is of interest to investigate whether sedentary behavior (SB), light physical activity (LPA) and moderate and vigorous physical activity (MVPA) are independently related to health outcomes. However, these variables often have high collinearity and the discretization might lead to loss of information. We consider functional and compositional data approaches for analyzing such data. The proposed approach is flexible and robust, unless existing models that rely on restrictive assumptions. We applied the proposed approach to an ancillary study of the Women's Health Initiative to study the association between physical activity and cardiometabolic biomarkers in 6,500 older women.

email: cdi@fredhutch.org

THE TENSOR MIXTURE MODEL FOR COMPOSITIONAL DATA WITH ESSENTIAL ZEROS: JOINTLY PROFILING ACCELEROMETRY-ASSESSED PHYSICAL ACTIVITY AND SEDENTARY BEHAVIOR IN THE HISPANIC COMMUNITY HEALTH STUDY / STUDY OF LATINOS (HCHS/SOL)

Daniela Sotres-Alvarez*, University of North Carolina, Chapel Hill
Angel D. Davalos, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Kelly Evenson, University of North Carolina, Chapel Hill
Amy H. Herring, Duke University

Accelerometer data is assessed as time spent in multicomponent behaviors of incremental intensity, which implies the sum of all components is constrained by the observation period. Jointly patterning accelerometry-assessed physical activity and sedentary behavior via latent class analysis is challenging because of the compositional nature of the data and the absence of certain intensities (e.g. no vigorous activity). Current latent class methods cannot account for these challenges. Motivated by profiling physical activity and sedentary behavior, we develop a Bayesian tensor mixture of product kernels model for modeling the joint distribution of compositional multivariate proportions with essential zeros. Our novel modeling construction makes the most efficient use of all the data and accounts for all essential zero configurations by: 1) leveraging the additive log ratio with a zero-inflated component, and 2) using the tensor probability specification to include all zero patterns in one unified model. We apply our method on a subset of accelerometer data from the HCHS/SOL (N ~ 3,000), and jointly profile weekday and weekend physical activity and sedentary behavior.

email: dsotres@email.unc.edu

FUNCTIONAL REGRESSION ON ACCELEROMETRY DATA IN THE NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

Elizabeth J. McGuffey*, United States Naval Academy
Ekaterina Smirnova, Virginia Commonwealth University
Andrew Leroux, Johns Hopkins University
Elham Mokhtari, University of Montana
Vadim Zipunnikov, Johns Hopkins University
Ciprian Crainiceanu, Johns Hopkins University

The advancement of accelerometer technology and the availability of large datasets reporting participants' minute-to-minute physical activity levels have facilitated investigations into associations between activity patterns and various health outcomes. One popular approach for including activity levels in regression models is to first summarize each participant's daily activity with a small number of scalar measures and subsequently include these as predictors in the model. In this presentation we explore an alternative approach: incorporating activity patterns as a functional predictor. In particular, we utilize a functional Cox proportional hazards model to analyze the association between daily activity patterns and mortality for a subset of participants from the 2003-2004 and 2005-2006 NHANES cohorts (n>3,200), and we discuss the potential benefits and drawbacks of the approach.

email: elizabeth.mcguffey@gmail.com

29. CAUSAL INFERENCE WITH NON-IGNORABLE MISSING DATA: NEW DEVELOPMENTS IN IDENTIFICATION AND ESTIMATION

IDENTIFICATION AND ESTIMATION OF CAUSAL EFFECTS WITH CONFOUNDERS MISSING NOT AT RANDOM

Shu Yang*, North Carolina State University
Linbo Wang, University of Toronto
Peng Ding, University of California, Berkeley

It is of great importance to draw causal inference from unconfounded observational studies, which, however, becomes challenging if the confounders are subject to missingness. Generally, causal effects are not identifiable if the confounders are missing not at random. In this paper, we propose a novel framework to nonparametrically identify the causal effects with confounders subject to an outcome-independent missingness, that is, the missing data mechanism is independent of the outcome, given the treatment and possibly missing confounders. We then propose nonparametric two-stage least squares estimation and parametric estimation for the average causal effect.

email: syang24@ncsu.edu

USING MISSING TYPES TO IMPROVE PARTIAL IDENTIFICATION WITH MISSING BINARY OUTCOMES

Zhichao Jiang*, Harvard University
Peng Ding, University of California, Berkeley

Frequently, empirical studies are plagued with missing data. When the data are missing not at random, the parameter of interest is not identifiable in general. Without imposing additional assumptions, we can derive bounds of the parameters of interest, which, unfortunately, are often too wide to be informative. Therefore, it is of great importance to sharpen these worst-case bounds by exploiting additional information. Traditional missing data analysis uses only the information of the binary missing data indicator, that is, a certain data point is either missing or not. Nevertheless, real data often provide more information than a binary missing data indicator, and they often record different types of missingness. Some missing types are more likely to be missing not at random, but other missing types are more likely to be missing at random. We show that making full use of the missing types allows us to obtain narrower bounds of the parameters of interest. In a real-life example, we demonstrate substantial improvement of more than 50\% reduction in bound widths for estimating the prevalence of HIV in rural Malawi.

email: zhichaoj89@gmail.com

IMPROVED EVALUATION OF HIV PREVALENCE ADJUSTING FOR INFORMATIVE NON-PARTICIPATION

Linbo Wang*, University of Toronto
Eric Tchetgen Tchetgen, University of Pennsylvania
Kathleen Wirth, Harvard T.H. Chan School of Public Health

HIV prevalence is routinely estimated from household surveys. Estimates may be biased if HIV-infected persons are less likely to participate. A design-based method leveraging interviewer features as instrumental variables (IV) for non-ignorable missing data can be used to account for such bias. Using an IV, we examined the presence and magnitude of bias due to missing data on HIV status within a large randomized trial in Botswana. Our findings suggest that HIV prevalence estimates which ignore non-participation may be downwardly biased. Investigators should consider including IVs in the study design to safeguard against non-participationinduced bias.

email: linbo.wang@utoronto.ca

BAYESIAN SPATIAL PROPENSITY SCORE ANALYSIS: UNMEASURED AND GEOGRAPHIC CONFOUNDING

Joon Jin Song*, Baylor University
Yawen Guan, North Carolina State University
Veronica Berrocal, University of Michigan
Bo Li, University of Illinois
Shu Yang, North Carolina State University

Observational data are increasingly being used for causal inference in social science and public health studies. Due to the lack of randomization, reliable causal inference cannot be made directly in observational studies. Propensity score analysis allows to carry out causal inference in observational studies and has become increasingly popular. The propensity score adjusts for exposure effects in the presence of confounding among different groups and removes any bias due to observed covariates if no unmeasured confounding exists. In spatially referenced observational data, geographic confounding might occur and play a role in causal inference. The objective of this study is to develop rigorous causal inference methods in the presence of both geographic and individual-level confounding. To address these issues in spatial observational data, we propose Bayesian propensity score methods using spatial random effects to account for the unmeasured confounding due to geographic proximity. A simulation study is performed to examine the impact of geographic confounding on treatment effect estimation. We also apply our proposed methods to a real-world dataset.

email: Joon_Song@baylor.edu

30. ADVANCED DEVELOPMENT IN JOINT MODELING AND RISK PREDICTION

DYNAMIC PREDICTION OF COMPETING RISK EVENTS USING LANDMARK SUB-DISTRIBUTION HAZARD MODEL WITH MULTIVARIATE LONGITUDINAL BIOMARKERS

Liang Li*, University of Texas MD Anderson Cancer Center

The cause-specific cumulative incidence function (CIF) quantifies the subject-specific disease risk with competing risks. With longitudinally collected biomarker data, it is of interest to dynamically update the predicted CIF by incorporating the most recent biomarker as well as the cumulating longitudinal history. Motivated by a longitudinal cohort study of chronic kidney disease, we propose a model dynamic prediction of end stage renal disease using multivariate longitudinal biomarkers, accounting for the competing risk of death. The proposed framework extends the landmark survival modeling to the competing risks data, and implies a distinct sub-distribution hazard regression model defined at each landmark time. The model parameters, prediction horizon, longitudinal history and at-risk population are allowed to vary over the landmark time. Local polynomial is used to estimate the model parameters without explicitly modeling its longitudinal trajectory. We conducted simulations to evaluate the performance of the estimation procedure and predictive accuracy. The methodology is illustrated with data from the African American Study of Kidney Disease and Hypertension.

email: LLi15@mdanderson.org

ESTIMATION UNDER COVARIATE-INDUCED DEPENDENT TRUNCATION THROUGH INVERSE PROBABILITY OF TRUNCATION WEIGHTING

Jing Qian*, University of Massachusetts
Bella Vakulenko-Lagun, Harvard T.H. Chan School of Public Health
Sy Han Chiou, University of Texas, Dallas
Rebecca A. Betensky, New York University

Many observational cohort studies are assembled using complex sampling schemes involving truncation. Ignoring the issue of truncation or incorrectly assuming quasi-independence can lead to study bias and incorrect interpretations. Available approaches to handle dependently truncated data are sparse and incomplete. We

propose an inverse probability weighting framework for estimating the survival function of a failure time subject to left truncation and possible right censoring. We develop several Kaplan-Meier type estimators based on inverse probability of truncation weighting. The proposed methods allow adjusting for informative truncation due to variables associated with both event time and truncation time. We conduct extensive simulation studies under various scenarios, including the sensitivity analysis under model misspecification. Simulation studies show that the proposed methods perform well in finite sample, and demonstrate the importance of model checking. We apply the proposed methods to a clinical study.

email: qian@umass.edu

JOINT MODELING OF MULTIPLE TIME-TO-EVENT OUTCOMES

Shanshan Zhao*, National Institute of Environmental Health Sciences, National Institutes of Health
Ross L. Prentice, Fred Hutchinson Cancer Research Center

Multivariate failure time data are important in biomedical research. While statistical methods for univariate failure time data are well established, the corresponding standard analysis tools for multivariate failure time data have not yet been established. The main difficulty is that with multiple censored time-to-disease outcomes, the joint likelihood is non-uniquely due to uninformative data points concerning the local dependency between event times. This talk will focus on some recent development in this area, including nonparametric estimates of joint survival function, average dependency measure, and semiparametric regression models of the cross ratio process. The proposed methods has the ability to explore and estimate dependency between event times as well as to understand the relationship between dependency and risk factors. Simulation evaluations as well as an application to the Women's Health Initiative's hormone therapy trial will be presented.

email: shanshan.zhao@nih.gov

PREDICTIVE ACCURACY OF SURVIVAL REGRESSION MODELS SUBJECT TO NON-INDEPENDENT CENSORING

Ming Wang*, The Pennsylvania State University
Qi Long, University of Pennsylvania
Chixiang Chen, The Pennsylvania State University

Survival regressions are commonly applied in clinical trials and biomedical studies, and evaluating their predictive performance plays an important role for model diagnosis and selection. The presence of censored data, particularly if informative, may post more challenge for assessment of predictive accuracy. Existing work mainly focus on prediction for survival probabilities but few for survival time. In this work, we will modify the original metric of mean square errors adjusted for the censoring mechanism of coarsening at random (CAR) or noncoarsening at random by adopting the technique of inverse probability of censoring weighting (IPCW). Our predictive metric can be adaptive to various survival regression frameworks including but not limited to accelerated failure time models. Also, we provide theoretical proof on the asymptotic properties of the IPCW estimators under CAR. To be specific, we consider both settings of low- and high-dimensional data to extend the proposed method and evaluate its the performance. Extensive simulation studies are conducted, and finally, the data from the Critical Assessment of Microarray Data Analysis is used for a further illustration.

email: mwang@phs.psu.edu

31. CLASSIFICATION AND VARIABLE SELECTION UNDER ASYMMETRIC LOSS

AN UMBRELLA ALGORITHM TO NEYMAN-PEARSON CLASSIFICATION

Xin Tong*, University of Southern California
Jingyi Jessica Li, University of California, Los Angeles
Yang Feng, Columbia University

In many binary classification applications, such as disease diagnosis and spam detection, practitioners commonly face the need to limit type I error (that is, the conditional probability of misclassifying a class 0 observation as class 1) so that it remains below a desired threshold. To address this need, the Neyman-Pearson (NP) classification paradigm is a natural choice; it minimizes type II error (that is, the conditional probability of misclassifying a class 1 observation as class 0) while enforcing an upper bound, alpha, on the type I error. Although the NP paradigm has a century-long history in hypothesis testing, it has not been well recognized and implemented in classification schemes. Common practices that directly limit the empirical type I error to no more than alpha do not satisfy the type I error control objective because the resulting classifiers are still likely to have type I errors much larger than alpha. This talk introduces an umbrella algorithm that adapts scoring-type classification methods to the Neyman-Pearson paradigm.

email: xint@marshall.usc.edu

AN UMBRELLA ALGORITHM THAT LINKS COST-SENSITIVE LEARNING TO NEYMAN-PEARSON CLASSIFICATION

Wei Vivian Li*, University of California, Los Angeles
Xin Tong, University of Southern California
Jingyi Jessica Li, University of California, Los Angeles

Cost-sensitive (CS) classification methods account for asymmetric misclassification costs and are widely applied in real-world problems such as medical diagnosis, transaction monitoring, and fraud detection. The current approaches to binary CS learning usually assign different weights to the two classes in non-unified ways, and the three main ways include rebalancing the sample before training, changing the objective function for training, and adjusting the estimated posterior class probabilities after training. Moreover, existing CS learning work has only focused on improving empirical classification errors or costs incurred by the assigned weights while overlooking the changes in population classification errors. We propose an umbrella algorithm to estimate the population type I error control achieved by multiple binary CS learning approaches. Our algorithm for the first time establishes a connection between CS learning and the Neyman-Pearson classification paradigm, which minimizes the population type II error while enforcing an upper bound on the population type I error.

email: liw@ucla.edu

NEYMAN-PEARSON CLASSIFICATION UNDER LABEL NOISE

Bradley Rava*, University of Southern California
Shunan Yao, University of Southern California

Label noise (i.e., imperfect labels) is a common problem occurred in industry settings. This creates a challenge for accurate classification, particularly when the proportion of mislabels are unknown.This work generalizes the NP umbrella algorithm to the general setting that has an unknown level of label noise.

email: brava@marshall.usc.edu

NEYMAN-PEARSON CRITERION (NPC): A BUDGET CONSTRAINED MODEL SELECTION CRITERION FOR ASYMMETRIC PREDICTION

Jingyi Jessica Li*, University of California, Los Angeles

Under the Neyman-Pearson (NP) classification paradigm for asymmetric classification, where the two types of classification errors do not have the same priority, we propose a model ranking criterion, Neyman-Pearson Criterion (NPC). NPC allows users to compare different feature sets in asymmetric binary prediction problems, and hence is useful for designing a predictive model for disease diagnosis. Specifically, NP classifiers (e.g., an NP logistic regression classifier) based on candidate feature sets will have the type I error (the more severe type of error) controlled under a user-specified threshold (e.g., .05) with high probability. NPC of each NP classifier is defined as the estimated type II error (the less severe type of classification error). Then the model with the smallest value of NPC will be chosen. Theoretical properties of NPC are derived. Extensive simulation studies show that NPC outperforms existing feature selection criteria when practitioners have a priority on the type I error. Real data applications to DNA methylation profiles of breast cancer patients and gene expression profiles of brain tumor patients further demonstrate the use and advantages of NPC.

email: jli@stat.ucla.edu

32. SPEED POSTERS: HIGH-DIMENSIONAL DATA/OMICS

32a. INVITED SPEED POSTER: BETTER DIAGNOSTIC AND PROGNOSTIC TOOLS FOR MULTIPLE SCLEROSIS BASED ON MRI

Russell T. Shinohara*, University of Pennsylvania

Multiple Sclerosis (MS) is an immune-mediated disease of the central nervous system whose hallmark is structural changes in the brain. Today, these changes are measured using MRI for diagnostic and disease-monitoring purposes. However, these changes can be concurrent or confused with other similar changes that occur, especially in older people. In this work, we study new statistical approaches for analyzing MRIs to provide biologically meaningful information about lesion structure, texture, shape, tissue damage, and spatial distribution with respect to brain vasculature. Taken together, these new computerized biomarkers promise improved diagnostic accuracy and more detailed prognostic information for clinical applications.

email: rshi@pennmedicine.upenn.edu

32b. INVITED SPEED POSTER: GROUP AND INDIVIDUAL NON-GAUSSIAN COMPONENT ANALYSIS FOR MULTI-SUBJECT fMRI

Benjamin B. Risk*, Emory University
Yuxuan Zhao, Cornell University
David S. Matteson, Cornell University

Independent component analysis (ICA) is an unsupervised learning method popular in functional magnetic resonance imaging. Group ICA has been used to identify biomarkers in neurological disorders including autism spectrum disorder and dementia. However, current methods use a PCA step that may remove low-variance features. Linear non-Gaussian component analysis (LNGCA) enables dimension reduction and component estimation simultaneously in single-subject fMRI. We present a group and individual LNGCA model to extract group components shared by more than one subject, as well as individual components unique to each subject. To determine the total number of components, we propose a parametric bootstrap that accounts for spatial dependence. In simulations, our estimated group components achieve higher accuracy compared to group ICA. Moreover, we recover the individual components for each subject. We apply our method to an fMRI study where the group signals include resting-state networks and individual components include artifacts unrelated to neuronal signal. The decomposition into group and individual components is a promising direction for feature detection in neuroimaging.

email: benjamin.risk@emory.edu

32c. REGULARIZED PREDICTION MODELING IN SMALL SAMPLES WITH APPLICATION TO PREDICTING TOXICITY IN A CAR T-CELL IMMUNOTHERAPY TRIAL

Mackenzie J. Edmondson*, University of Pennsylvania
David T. Teachey, Children's Hospital of Philadelphia
Pamela A. Shaw, University of Pennsylvania

Novel cancer therapies have increased the need for early phase clinical trials to determine the efficacy and safety of particular therapies for a given patient. In the age of personalized medicine, identifying biomarkers predictive of toxicity can improve patient care and reveal new therapeutic targets. Risk prediction models are often used to estimate risk of toxicity given certain biomarker levels. Approaches for constructing these models have mostly been evaluated for large samples; in early phase clinical trials, however, samples are typically small, creating a need for assessment of these methods in a small-sample setting. Through structured simulations, we compare the performance of several regularized logistic regression procedures in a high-dimensional, small-sample setting: lasso, elastic net, and ridge regression, along with a commonly-used stepwise regression technique. Simulations were modeled after a real data setting with highly correlated biomarkers and a relatively low sample size. We apply these methods to guide selection of an approach to predict risk of cytokine release syndrome following treatment with CAR T-cell immunotherapy in a small pediatric cohort.

email: macjohn@pennmedicine.upenn.edu

32d. BAYESIAN GWAS WITH STRUCTURED AND NON-LOCAL PRIORS

Adam Kaplan*, University of Minnesota
Eric F. Lock, University of Minnesota
Mark Fiecas, University of Minnesota

We introduce a novel Bayesian approach to genome-wide association studies (GWAS) that improves over existing methods in two important ways. First, we describe a model that allows for marker characteristics to influence its probability of association with an outcome. For this we use a hierarchical Dirichlet Process (DP) model that allows for clustering of the genes in tandem with a regression model for marker-level covariates. Second, we use Non-Local priors to model the difference in probability of minor allele status between phenotype status. We outline the implementation of and discuss the philosophical problems treated by Non-Local priors within the GWA framework. One problem is that current Bayesian model comparison methods have asymptotic rates of convergence favoring the alternative hypothesis over the null, whereas we define a Non-Local prior for the GWAS context that gives symmetric rates. We assess the components of our model with simulation studies under several different scenarios. We apply our method to SNP data collected from Alzheimer's disease and cognitively normal patients from the Alzheimer's Database Neuroimaging Initiative.

email: kapla271@umn.edu

32e. INTERPRETABLE ADVANCE-LEARNING FOR DERIVING OPTIMAL DYNAMIC TREATMENT REGIMES WITH OBSERVATIONAL DATA

Aaron M. Sonabend*, Harvard T.H. Chan School of Public Health
Tianxi Cai, Harvard T.H. Chan School of Public Health
Peter Szolovits, MIT Computer Science and Artificial Intelligence Laboratory

Dynamic treatment regimes (DTRs) allow for the incorporation of patient heterogeneity in making optimal sequential treatment decisions, pushing personalized medicine forward. To this end, flexible statistical and machine learning methods are being increasingly employed in hospitals and used by clinicians. In this context, there is a critical need for models that allow for interpretability, accommodate causal frameworks, and can be trained from observational data. We propose an Advance-learning method that is both flexible for learning the complex structure of the data, and lends itself to intuitive interpretations for the relationships of interest. To capture high-dimensional complex structures we modeled the Q function for baseline treatment using deep neural networks. To allow for multivariate treatments and interaction effects on outcome, we modeled the contrast function as a multinomial log-linear model. In this project we derive the method and show the estimating equations and their large sample properties. Additionally, an application for finding optimal DTRs, and interpreting causal relationships is illustrated using a Sepsis cohort from the MIMIC-III database.

email: asonabend@g.harvard.com

32f. OMNIBUS WEIGHTING INCORPORATING MULTIPLE FUNCTIONAL ANNOTATIONS FOR WHOLE GENOME SEQUENCING RARE VARIANT ASSOCIATION STUDIES

Xihao Li*, Harvard University
Zilin Li, Harvard University
Hufeng Zhou, Harvard University
Yaowu Liu, Harvard University
Han Chen, University of Texas School of Public Health at Houston
Alanna C. Morrison, University of Texas School of Public Health at Houston
Eric Boerwinkle, University of Texas School of Public Health at Houston
Xihong Lin, Harvard University

Whole-genome sequencing studies have been increasingly conducted to investigate associations between susceptible rare variants and traits. Existing variant-set tests only incorporate MAF as weights in rare variant association analysis. External biological information based on various functional annotations can be utilized to predict variant functionality, and thus help boost power. In this paper, we develop variant-Set Test for Association using Annotation infoRmation (STAAR), a general framework that incorporates multiple functional annotations as an omnibus weighting scheme to boost power for variant-set tests in WGS studies. We propose using Annotation Principle Components, which perform multi-dimensional summary of function annotations, to capture multiple aspects of biological functionality. We derive three tests in the STAAR framework by incorporating multiple annotations, including STAAR-B/S/O. Simulation studies show that the proposed STAAR tests control type I error rates and achieve greater power compared to conventional tests. We also illustrate the proposed framework to a WGS analysis from the Atherosclerosis Risk in Communities study for lipid traits.

email: xihaoli@g.harvard.edu

32g. DETECTING GENES WITH ABNORMAL CORRELATION AMONG METHYLATION SITES

Hongyan Xu*, Augusta University

Abberant DNA methylation has been involved in cancer development. Current approaches focus on detecting mean and/or variences differences between different tumors or tumors and normal controls. In this study, we examined the pairwise correlations between CpG sites in colorectal cancer samples and normal controls. Mutiple genes with abnormal correlations between CpG sites have been identified when the correlations were compared between the tumor and normal controls. We replicated the findings in an independent colorectal cancer sample. Our approach deteced more genes with disruption of pairwise methylation correlation than the approach for mean/variance differences. This provides a new approach for identifying genes related to the etiology of colorectal cancer and is readily applicable for similar studies in other tumor types.

email: hxu@augusta.edu

32h. LETTING THE LaxKAT OUT OF THE BAG: PACKAGING, SIMULATION, AND NEUROIMAGING DATA ANALYSIS FOR A POWERFUL KERNEL TEST

Jeremy S. Rubin*, University of Maryland, Baltimore County
Simon Vandekar, Vanderbilt University
Lior Rennert, Clemson University
Mackenzie Edmonson, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

Biomedical research areas including genomics and neuroimaging often have a number of independent variables that is much greater than the sample size. The sequence kernel association test (KAT) and Sum of Scores tests can offer improved power in this feature setting; however, power is significantly reduced in the presence of a large number of unassociated independent variables. We propose the linear Maximal KAT (LaxKAT), which has maximizes the KAT test statistic over a subspace of linear kernels to increase power. A permutation testing scheme is used to estimate the null distribution of the LaxKAT statistic and perform hypothesis testing. Calculation of the LaxKAT was implemented in R with a modular design to allow for greater usability. We find that this test is able to reliably detect both simple and complex signals for two simulation studies. It is expected that the LaxKAT will have competitive power relative to the projected score test when applied to detect predictors of memory impairment in cortical thickness measurements from the Alzheimer's Disease Neuroimaging Initiative study.

email: jeremy22114@gmail.com

32i. FUNCTIONAL DATA ANALYSIS FOR MAGNETIC RESONANCE SPECTROSCOPY (MRS) DATA IN SPINOCEREBELLAR ATAXIAS

Lynn E. Eberly*, University of Minnesota
Meng Yao, University of Minnesota
James Joers, University of Minnesota
Gulin Oz, University of Minnesota

Functional data analysis was applied to magnetic resonance spectroscopy (MRS) data to classify spinocerebellar ataxias (SCAs) vs. controls, and to predict the Scale for the Assessment and Rating of Ataxia (SARA) and Activities of Daily Living (ADL) for SCA participants. 7T MRS was acquired from the pons of 68 SCAs and 18 controls. The performance of full spectrum data for classification and prediction was compared to that of quantified concentrations of specific metabolites. Metabolite concentrations resulted in better classification performance, quantified by leave one out cross validated (LOOCV) misclassification rate, compared to using the full spectrum, but the rate varied by SCA type (SCA 1, 2, 3, and 6). Full spectrum data and metabolite concentrations had similar LOOCV mean squared error (MSE) performance for score predictions. Functional data analysis of MRS spectra has the potential to improve disease classification and prediction of disease severity and quality of life measures, but that was not realized here. However, the influence of the many signal processing steps on the final spectra, to ensure comparability across participants, needs to be better understood.

email: leberly@umn.edu

32j. A LOCAL TEST FOR GROUP DIFFERENCES IN SUBJECT-LEVEL MULTIVARIATE DENSITY DATA

Jordan D. Dworkin*, University of Pennsylvania
Russell T. Shinohara, University of Pennsylvania

In neuroimaging research, magnetic resonance imaging (MRI) modalities yield multiple intensity values at each voxel. Currently, voxel-level studies rarely incorporate multi-modal information, and often assume that effects occur in the same voxels across subjects. The current study decouples voxel intensities from their physical location, and operationalizes MRI data as subject-level multivariate voxel densities. Within this framework, a test for group differences at each location within multivariate distributions is developed, and its asymptotic properties are presented. This method is applied to T1 and FLAIR scans in subjects with multiple sclerosis. Joint T1/FLAIR densities were compared between relapsing-remitting (n=32) and secondary-progressive (n=23) subjects, revealing a significant difference in the prevalence of voxels with high intensity on both modalities ($p<0.01$). This effect was not found in the means or standard deviations within either modality, nor was it reflected in the correlation between the two modalities. This method represents a step forward in utilizing joint information in MRI data to discover disease-related voxel intensity profiles.

email: jdwor@pennmedicine.upenn.edu

32k. SEMI-PARAMETRIC DIFFERENTIAL ABUNDANCE ANALYSIS FOR METABOLOMICS AND PROTEOMICS DATA

Yuntong Li*, University of Kentucky
Arnold J. Stromberg, University of Kentucky
Chi Wang, University of Kentucky
Li Chen, University of Kentucky

Identifying differentially abundant features between different experimental conditions is a common goal for many metabolomics and proteomics studies. However, analyzing metabolomics and proteomics data from mass spectrometry is challenging because the data may not be normally distributed and contain a large fraction of zero values. Although several statistical methods have been proposed, they either require data normality assumption, or are inefficient. We propose a new Semi-parametric Differential Abundance analysis (SDA) method for metabolomics and proteomics data from mass spectrometry. The method considers a two-part model, a logistic regression for the zero proportion and a semi-parametric log-linear model for the non-zero values. Our method is free of distributional assumption and also allows for adjustment of covariates. We propose a kernel-smoothed likelihood method to estimate regression coefficients in the two-part model and construct a likelihood ratio test for differential abundant analysis. Simulations and real data analyses demonstrate that our method outperforms existing methods.

email: yli362@g.uky.edu

32l. MIXnorm: NORMALIZING GENE EXPRESSION DATA FROM RNA SEQUENCING OF FORMALIN-FIXED PARAFFIN-EMBEDDING SAMPLES

Shen Yin*, Southern Methodist University and University of Texas Southwestern Medical Center

Xinlei Wang, Southern Methodist University
Gaoxiang Jia, Southern Methodist University and University of Texas Southwestern Medical Center
Yang Xie, University of Texas Southwestern Medical Center

Recent studies have shown that RNA sequencing (RNA-seq) can be used to measure mRNA of sufficient quality extracted from Formalin-Fixed Paraffin-Embedded (FFPE) tissues to provide whole-genome transcriptome analysis. However, little attention has been given to the normalization of such data, a key step that adjusts for unwanted biological and technical effects. Existing methods, developed based on fresh frozen or similar-type samples, may cause suboptimal performance. We proposed a new normalization method, labeled MIXnorm, for RNA-seq data from FFPE samples. MIXnorm relies on a two-component mixture model, which models non-expressed genes by zero-inflated Poisson distributions and models expressed genes by truncated Normal distributions. To obtain the maximum likelihood estimates, we developed a nested EM algorithm, in which closed-form updates are available in each iteration. By eliminating the need of Monte Carlo simulation in the E-step, the algorithm is easy to implement and computationally efficient. We evaluated MIXnorm through simulations and two cancer studies. MIXnorm makes significant improvement over commonly used methods for RNA-seq expression data.

email: yinshen1992@live.com

33. PERSONALIZED MEDICINE

ESTIMATING INDIVIDUALIZED TREATMENT REGIMES FROM CROSSOVER STUDIES

Crystal T. Nguyen*, University of North Carolina, Chapel Hill
Daniel J. Luckett, University of North Carolina, Chapel Hill
Grace E. Shearrer, University of North Carolina, Chapel Hill
Anna R. Kahkoska, University of North Carolina, Chapel Hill
Donna Spruijt-Metz, University of Southern California
Jaimie N. Davis, University of Texas, Austin
Michael R. Kosorok, University of North Carolina, Chapel Hill

Precision medicine aims to tailor treatment to each individual when diseases present with heterogeneity across patients. It is of particular interest to estimate an optimal individualized treatment regime (ITR) that recommends decisions based on patient characteristics to maximize the mean of some outcome. There exist several methods for estimating an optimal ITR from clinical trial data in trials where each subject is assigned to a single treatment. However, little work has been done for crossover study designs which naturally lend themselves to precision medicine, because they allow for observing the response to multiple treatments for each patient. In this paper, we introduce a method to estimate the optimal ITR using data from a 2x2 crossover study with carryover effects. The proposed method is similar to policy search methods; however, we take advantage of the study design by considering the difference in response as the observed reward. We establish Fisher and global consistency, present numerical experiments, and analyze data from a feeding trial using the proposed method to demonstrate its improved performance compared to standard methods for a parallel study design.

email: cnguyen6292@gmail.com

BIOMARKER SCREENING IN THE LEARNING OF INDIVIDUALIZED TREATMENT RULES VIA NET BENEFIT INDEX

Yiwang Zhou*, University of Michigan
Haoda Fu, Eli Lilly and Company
Peter X.K. Song, University of Michigan

One central task of personalized medicine is to establish individualized treatment rules (ITRs) for patients with heterogeneous responses to different treatments. Motivated from a diabetes clinical trial, we consider a problem where many biomarkers are potentially useful to improve an existing ITR. This calls for a screening procedure to assess the added values of new biomarkers to derive an improved ITR. We propose a new test based on net benefit index (NBI) that quantifies gain or loss of treatment benefit due to reclassification in which the optimal labels are obtained by support vector machine (SVM) in the context of outcome weighted learning (OWL). We calculate the p-value of the proposed NBI-based test using the bootstrap null distribution generated by stratified permutations on individual treatment arm. The performance of the proposed method is evaluated by simulations and the motivating clinical trial. Our results show that the NBI-based test controls the false discovery rate well and achieves high sensitivity. In addition, this screening method demonstrates an improved correct classification rate when ITR is expanded by including selected biomarkers.

email: yiwangz@umich.edu

INTEGRATIVE LEARNING TO COMBINE INDIVIDUALIZED TREATMENT RULES FROM MULTIPLE RANDOMIZED TRIALS

Xin Qiu*, Janssen Research & Development
Donglin Zeng, University of North Carolina, Chapel Hill
Yuanjia Wang, Columbia University

Implementing individualized treatment rules (ITRs) that adapt to patient's characteristics and intermediate responses holds promise to improve treatment response of mental disorders. However, several barriers, in particular, lack of power to detect treatment modifiers as tailoring variables and lack of generalizability or reproducibility of ITRs derived from a single study, pose significant challenges to clinical practice. In this work, we propose a novel integrative learning method to combine evidence from multiple clinical trials to yield an integrative ITR that improves both precision and reproducibility. Our method does not require all studies to use the same set of covariates, and thus allows study-specific ITRs to be transferable across studies. Specifically, to transfer information we propose integrative learning to enhance a high-resolution ITR by borrowing information from coarsened ITRs or vice versa. We demonstrate that the integrative ITR yields a greater benefit and has improved precision compared to single-study ITRs or non-personalized rules by extensive simulation studies and an application to multiple clinical trials of major depressive disorder.

email: qiuxin1026@gmail.com

DISPARITY SUBTYPING: BRINGING PRECISION MEDICINE CLOSER TO DISPARITY SCIENCE

Huilin Yu*, University of Miami
J. Sunil Rao, University of Miami
Jean Eudes Dazard, Case Western Reserve University

Disparities researchers have begun looking to the precision medicine paradigm with the hope that some incorporation of its principles will allow for a more focused and precise path forward to reduce population disparities. While the emphasis may switch to populations from individuals, central to the paradigm still is the ability to classify individuals into subpopulations who differ in meaningful ways with respect to underlying biology and outcomes. How to do such a thing in disparity science has been proven elusive since it requires identifying disparity subpopulations which is a somewhat abstract concept. In this paper we present two different strategies - level set estimation and peeling. Both are based on a recursive partitioning algorithm. The former is combined with clustering similar partitions; the latter adopts a strategy of sequentially searching and then extracting extreme difference subgroups in a population. Using series of simulation studies and then an analysis of ovarian cancer survival in patients from The Cancer Genome Atlas (TCGA) repository, we demonstrate that such disparity subtypes can indeed be found, characterized, and then validated on test data.

email: yuhuilin619@gmail.com

MODIFIED THOMPSON SAMPLING FOR PRECISION MEDICINE

John Sperger*, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

A central task in precision medicine is to determine a treatment rule for assigning treatments to patients based on baseline characteristics and biomarker measurements, and the question of how to design experiments when the goal is treatment rule discovery is an important open question. A key tension in patient allocation for clinical trials is balancing the task of gaining as much information as possible to reduce out-of-sample errors against the goal of reducing the number of patients assigned inferior treatments. To address this, we investigate two modified Thompson sampling algorithms which randomize patients using the posterior predictive distribution of success under each treatment given their baseline characteristics and an additional random component whose purpose is to improve information gain. Using simulation studies, we compare the effectiveness of these randomization schemes and traditional methods in terms of relative efficiency for treatment rule estimation, the number of patients assigned to suboptimal treatments, and the out-of-sample classification error for future patients.

email: jsperger@live.unc.edu

ESTIMATING INDIVIDUALIZED DECISION RULES WITH TAIL CONTROLS

Zhengling Qi*, University of North Carolina, Chapel Hill
Jong-Shi Pang, University of Southern California
Yufeng Liu; University of North Carolina, Chapel Hill

With the emergence of precision medicine, estimating optimal individualized decision rules (IDRs) has attracted tremendous attentions in many scientific areas. Most existing literature has focused on finding optimal IDRs that can maximize the expected outcome for each individual. Motivated by complex individualized decision making procedures and popular conditional value at risk (CVaR) measure, we propose two new robust criteria to estimate optimal IDRs: one is to control the average lower tail of the subjects' outcomes and the other is to control the individualized lower tail of each subject's outcome. Besides optimizing the individualized expected outcome, our proposed criteria take risks into consideration, and thus the resulting IDRs can prevent adverse events caused by the heavy lower tail of the outcome distribution. From the perspective of duality theory, the optimal IDR under our criteria can be interpreted as the decision rule that maximizes the worst-case scenario of the individualized outcome within a probability ambiguity set. Simulation studies and a real data application are used to further demonstrate the robust performance of our methods.

email: qizl1027@live.unc.edu

34. EPIDEMIOLOGIC METHODS

BAYESIAN PIECEWISE LINEAR MIXED MODELS WITH A RANDOM CHANGE POINT: AN APPLICATION TO STUDY EARLY GROWTH PATTERNS IN THE DEVELOPMENT OF TYPE 1 DIABETES

Xiang Liu*, The TEDDY Study Group, University of South Florida
Yangxin Huang, The TEDDY Study Group, University of South Florida
Kendra Vehik, The TEDDY Study Group, University of South Florida
Jeffrey Krischer, The TEDDY Study Group, University of South Florida

To study the pathogenesis of Type 1 diabetes (T1D), proposed accelerator and overload hypotheses postulate that overweight and rapid growth speed up both beta cell insufficiency and an increased insulin resistance. A child's growth (weight) trajectory during childhood starts with a phase of fast growth and then a phase of slow growth. An individual's timing of growth change is important because it might be associated with the risk for either 1) islet autoimmunity, 2) clinical onset of T1D, or both. Here, we introduce a Bayesian two-phase piecewise linear mixed model, where the "change point" is an individual-level random effect corresponding to the timing connecting the two growth phases. This method is used to estimate the weight trajectories for children from the Environmental Determinants of Diabetes in the Young (TEDDY) study and then assess the association between the random effects (intercept, pre-change slope, post-change slope, change point) and the risk of either 1) islet autoimmunity or 2) clinical onset. The pre-change slope was significantly associated with the risk of islet autoimmunity.

email: xiang.liu@epi.usf.edu

A SPATIAL BAYESIAN HIERARCHICAL MODEL FOR COMBINING DATA FROM PASSIVE AND ACTIVE INFECTIOUS DISEASE SURVEILLANCE SYSTEMS

Xintong Li*, Emory University
Howard Chang, Emory University
Lance Waller, Emory University
Qu Cheng, University of California, Berkeley
Philip Collender, University of California, Berkeley
Justin Remais, University of California, Berkeley

Infectious disease surveillance data are important for monitoring disease burden and occurrence and informing efforts to protect and improve population health. Passive surveillance typically provides wide spatial coverage, but is subject to biases arising from differences in care-seeking behavior, diagnostic practices, and underreporting, which may vary in space and time. Active surveillance minimizes these biases, but is typically constrained to small areas and subpopulations due to resource limitations. Methods based on linkage of individual records between passive and active surveillance datasets provide a means to estimate and correct for the biases of either system, leveraging the size and coverage of passive surveillance and the quality of data in active surveillance. We developed a spatial Bayesian hierarchical model to estimate the sensitivity of passive and active surveillance for tuberculosis (TB) in Sichuan, China in 2010. We estimated that the active surveillance system has 80% (95% credible interval: 74%, 86%) sensitivity on average, while the average sensitivity of the passive system is 22% (95% CI: 17%, 43%).

email: garyli86@gmail.com

A JOINT MODEL OF OPIOID TREATMENT ADMISSIONS AND DEATHS FOR ADULTS AND ADOLESCENTS IN OHIO COUNTIES

David M. Kline*, The Ohio State University
Staci A. Hepler, Wake Forest University

Opioid misuse is a national epidemic and a significant public health issue due to its high prevalence of associated morbidity and mortality. Ohio has been hit as hard by the opioid epidemic as any state in the country. We are interested in characterizing rates for both adolescents and adults as they require differing treatment strategies and policy responses. We propose a joint spatial factor model that estimates county-level rates of opioid related treatment admission and death for adolescents and adults. By jointly modeling, we can borrow strength from the adult model to stabilize estimates for adolescents which are based on smaller counts. We will also estimate effects of county-level social environmental covariates to better characterize differences between counties. We will highlight gains in efficiency from use of this approach and discuss other statistical considerations for this application.

email: david.kline@osumc.edu

INFERENCE FOR CASE-CONTROL STUDIES INCLUDING PREVALENT CASES, AND PROSPECTIVE SURVIVAL INFORMATION

Soutrik Mandal*, National Cancer Institute, National Institutes of Health
Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Ruth Pfeiffer, National Cancer Institute, National Institutes of Health

In studies involving rare diseases like cancer, availability of incident cases may be limited. It is thus appealing to incorporate prevalent cases in such studies. However, not all prevalent cases survive until the beginning of the study and hence inclusion of only living prevalent cases causes bias. An extended version of the exponential tilting model can be used to correct such survival bias (Maziarz et al. 2018+). Sometimes, additional information on survival after disease onset is available, either for incident cases or both, incident and prevalent cases. We propose a two-stage method to estimate disease-exposure association when both incident and prevalent cases are included in a study along with their prospectively observed survival times. We derive the asymptotic properties and study the sensitivity of our method under different violations of assumptions and apply our method to a real dataset.

email: soutrikmandal@gmail.com

EFFECT SIZE MEASURES FOR MEDIATION ANALYSIS OF MULTIPLE CORRELATED EXPOSURES

Yue Jiang*, University of North Carolina, Chapel Hill
Shanshan Zhao, National Institute of Environmental Health Sciences, National Institutes of Health
Jason Peter Fine, University of North Carolina, Chapel Hill

Our research aim is to investigate the extent to which the relationship between smoking and lung function is mediated by DNA methylation, particularly by ranking mediation strength of nine cytosine-phosphate-guanine (CpG) sites previously identified as highly statistically significant. However, little existing research has focused on effect size. Furthermore, existing measures were developed solely for a single exposure, whereas accurately quantifying the effect of smoking requires consideration of the joint effect of multiple correlated variables. We propose a new effect size measure for multiple correlated exposures based on decomposition of explained variance. We further propose orthogonality constraints of effect estimates to address statistical and interpretational concerns. Closed forms and asymptotic properties are derived for all measures and supported by numerical simulation. Although all candidate CpG sites were highly significant mediators, cg03636183 (F2RL3), cg21566642 (ALPPL2), and cg05575921 (AHRR) were the strongest. These results allow for future biological research to be focused on CpG sites with the greatest mediational effects.

email: yuejiang@live.unc.edu

REGRESSION ANALYSIS OF COMBINED INCIDENT AND PREVALENT COHORT DATA

Chi Hyun Lee*, University of Massachusetts
Jing Ning, University of Texas MD Anderson Cancer Center
Richard Kryscio, University of Kentucky
Yu Shen, University of Texas MD Anderson Cancer Center

The Nun Study, a longitudinal study to examine risk factors for the progression of dementia, consists of subjects who were already diagnosed with dementia (i.e., prevalent cohort) and those who do not have dementia (i.e., incident cohort) at study enrollment. When assessing the risk factors' effects on the survival time from dementia diagnosis until death, utilizing data from both cohorts supports more efficient statistical inference because the two cohorts provide valuable complementary information. A major challenge in analyzing the combined cohort data is that the prevalent cases are not representative of the target population. Moreover, the dates of dementia diagnosis are not ascertained for the prevalent cohort in the Nun Study. Hence, the survival time for the prevalent cohort is only partially observed from study enrollment until death or censoring, with the time from dementia diagnosis to study enrollment missing. In this talk, I will discuss an efficient estimation method that uses both incident and prevalent cohorts under the proportional mean residual life model.

email: chihyunlee@umass.edu

35. STATISTICAL GENETICS: SINGLE-CELL SEQUENCING/ TRANSCRIPTOMIC DATA

SINGLE-CELL RNA SEQUENCING: NORMALIZATION FOR TECHNICAL NOISE AND BATCH EFFECTS

Nicholas J. Lytal*, University of Arizona
Di Ran, University of Arizona
Lingling An, University of Arizona

Gene sequencing experiments allow researchers to analyze the genetic content of cells in areas such as cancer and embryo studies. Single-cell RNA sequencing (scRNA-seq) has expanded the scope of this analysis beyond the limitations of bulk sequencing, allowing better identification of rare cell types and heterogeneity within cell groups. Even so, insufficient input material on a single-cell scale leads to concerns with amplification bias, as well as the potential for dropout events. Furthermore, batch effects are more pronounced in single-cell experiments, which frequently require separate sequencing runs for cell groups. To answer these concerns, several normalization methods have been developed that correct observed gene counts to better represent the actual counts present in a sample. We propose a scRNA-seq normalization method that employs normalization between identified cell groups and data imputation to correct for technical noise, dropout events, and batch effects. We compare this method with existing normalization methods using real data sets acquired from Illumina GAIIx and HiSeq sequencing platforms.

email: njlytal@email.arizona.edu

BULK GENE EXPRESSION DECONVOLUTION BY SINGLE-CELL RNA SEQUENCING

Meichen Dong*, University of North Carolina, Chapel Hill
Yuchao Jiang, University of North Carolina, Chapel Hill
Fei Zou, University of North Carolina, Chapel Hill

Single-cell RNA sequencing (scRNA-seq) enables characterization of transcriptomic profiles and circumvents averaging artifacts associated with traditional bulk RNA-seq data. We devise a framework that leverages scRNA-seq data to profile cell-type specific gene expressions and estimate cell-type compositions from bulk RNA-seq, which distinguishes existing methods in: scaling the raw single-cell read-count matrix by a gene- and donor-specific maximal variance weight so that residuals from genes with larger weights have smaller impact on cell-type composition estimation; removing misclassified cells instead of taking cell-type memberships as ground truth to improve robustness; an ensemble method to integrate deconvolution results across methods and datasets, addressing the batch-effect confounding when multiple scRNA-seq reference sets are available; adapting to the case-control setting and testing for differential cell-type composition and gene expression pattern. Our method is benchmarked against existing ones using pseudo-bulk samples generated in silico, whose true underlying cell type identities are known, and further applied to a real dataset as demonstration.

email: meichen@live.unc.edu

BIVARIATE ZERO-INFLATED NEGATIVE BINOMIAL (BZINB) MODEL FOR MEASURING DEPENDENCE

Hunyong Cho*, University of North Carolina, Chapel Hill
Di Wu, University of North Carolina, Chapel Hill

Zero inflated counts are frequently observed in single cell RNAseq and micorbiome sequencing data. When measuring dependence between count variables, nonparametric measures such as Pearson correlation and empirical mutual information often fail to exploit the distributional characteristics of count data. For example, single cell RNA sequencing data often contain high percentage of zero counts with a significant portion of large values at the same time. When many of the zeros are considered technical dropouts, this zero-inflation should not contribute toward measuring dependence. We propose a bivariate zero-inflated negative binomial model to better extract the dependence information from data. While existing bivariate negative binomial models are either overly complicated or not flexible enough, our model is constructed using layers of latent variables, providing rich but parsimonious distributional characteristics. In addition, it enables measuring the underlying dependence by adjusting for the zero-inflation component. Finally, EM-based estimation scheme, real data based simulation, and application to gene-set testing will be presented.

email: hunycho@live.unc.edu

SCOPE: A NORMALIZATION AND COPY NUMBER ESTIMATION METHOD FOR SINGLE-CELL DNA SEQUENCING

Rujin Wang*, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill
Yuchao Jiang, University of North Carolina, Chapel Hill

Whole-genome single-cell DNA sequencing (scDNA-seq) enables characterization of copy number profiles at the cellular level. ScDNA-seq data is, however, sparse, noisy, and highly variable even within a homogeneous cell population, due to biases and artifacts. We propose SCOPE, a normalization and copy number estimation method by scDNA-seq. The distinguishing features of SCOPE include: (i) utilization of cell-specific Gini coefficients to identify normal cells as negative control samples in a Poisson latent factor model; (ii) modeling of GC content bias using an EM algorithm embedded in the Poisson generalized linear models, which accounts for the different non-null copy number states; (iii) a cross-sample segmentation procedure to identify shared breakpoints across cells from the same genetic background. SCOPE outperforms existing methods for more accurate estimation of copy number aberrations and higher correlation with array-based copy number calls of purified bulk samples. We further demonstrate SCOPE on two recently released datasets using the 10X Genomics single-cell CNV pipeline and show that it can reliably recover 1% of the cancer cells from a background of normals.

email: rujin@email.unc.edu

MACAM: A SEMI-SUPERVISED STATISTICAL DECONVOLUTION METHOD FOR MIXED TRANSCRIPTOMIC DATA

Li Dong*, University of North Carolina, Chapel Hill
Fei Zou, University of North Carolina, Chapel Hill
Xiaojing Zheng, University of North Carolina, Chapel Hill

Deconvolution for heterogeneous samples is a very important step for differential expression analysis since tissue heterogeneity is a major confounding factor on differential gene expression analysis with bulk samples. Existing supervised deconvolution methods require prior information of known proportions, gene signatures or marker genes for each cell type, which are usually unavailable for many tissue cell types. Unsupervised methods try to solve this problem by completely disregarding markers information therefore loss of substantial accuracy. To fill in the methodological gap, we propose a marker assisted transcriptional heterogeneity deconvolution method, MACAM. The method extends unsupervised subpopulation maker genes identification method, convex analysis of mixtures (CAM), by incorporation of partial known markers in parallelism of latent variable model and the theory of convex sets to guide discovery of novel markers. Simulations demonstrate that MACAM has improved accuracy compared to CAM. Real data analyses are used to illustrate the performance of the proposed method.

email: doli@live.unc.edu

DETECTING REGULATORY GENETIC VARIANTS WITH TRANSCRIPTION FACTOR BINDING AFFINITY TESTING

Sunyoung Shin*, University of Texas, Dallas
Chandler Zuo, A.R.T. Advisors
Sunduz Keles, University of Wisconsin, Madison

Understanding the regulatory roles of non-coding genetic variants has become a central goal for interpreting results of genome-wide association studies. The regulatory significance of the variants may be interrogated by assessing their impact on transcription factor binding. We propose an efficient and scalable motif-based regulatory variant discovery tool, named atSNP (affinity testing for regulatory SNP detection). atSNP implements an importance sampling algorithm coupled with a first-order Markov model for the background nucleotide sequences to evaluate motif matches to both reference and variant alleles and assess variant-led changes in motif matches. Further, we have developed atSNP Search, a comprehensive web database for identifying human regulatory variants with statistical significance obtained from atSNP and composite logo plots, which are graphical representations of motif matches. atSNP Search users can test more than 37 billion variant-motif pairs with marginal significance in motif matches or alteration. Computational evidence from atSNP Search, when combined with experimental validation, may help with the discovery of disease mechanisms.

email: sunyoung.shin@utdallas.edu

36. MACHINE LEARNING AND TESTING WITH HIGH DIMENSIONAL DATA

INTEGRATIVE LINEAR DISCRIMINANT ANALYSIS WITH GUARANTEED ERROR RATE IMPROVEMENT

Quefeng Li*, University of North Carolina, Chapel Hill
Lexin Li, University of California, Berkeley

Multiple types of data measured on a common set of subjects arise in many areas. Numerous empirical studies have found that integrative analysis of such data can result in better statistical performance. However, the advantages of integrative analysis have mostly been demonstrated empirically. In the context of two-class classification, we propose an integrative linear discriminant analysis method, and establish a theoretical guarantee that it achieves a smaller classification error than running linear discriminant analysis on each data type individually. We also address the issue of outliers that is frequently encountered in integrative analysis.

email: quefeng@email.unc.edu

HIGH-DIMENSIONAL DECOMPOSITION-BASED CANONICAL CORRELATION ANALYSIS

Hai Shu*, University of Texas MD Anderson Cancer Center
Xiao Wang, Purdue University
Hongtu Zhu, University of North Carolina, Chapel Hill
Peng Wei, University of Texas MD Anderson Cancer Center

It is often seen in biomedical studies that multiple large-scale datasets are measured on a common set of objects. A typical model for jointly analyzing two such datasets is to decompose each data matrix into three parts: a low-rank common matrix that captures the shared information across datasets, a low-rank distinctive matrix that characterizes the individual information within a single dataset, and an additive noise matrix. Existing decomposition methods often focus on the orthogonality between the common and distinctive matrices, but inadequately consider the more necessary orthogonal relationship between the two distinctive matrices. The latter guarantees that no more shared information is extractable from the distinctive matrices. We propose decomposition-based canonical correlation analysis, a novel decomposition method that carefully constructs such orthogonality. The proposed estimators of common and distinctive matrices are shown to be consistent and have reasonably better performance than some state-of-the-art methods in both simulated data and the real data analysis of breast cancer genomic data.

email: shuhai.edu@gmail.com

ESTIMATION OF TUMOR IMMUNE CELL CONTENT USING SINGLE-CELL RNA-Seq DATA

Christopher M. Wilson*, H. Lee Moffitt Cancer Center
Xuefeng Wang, H. Lee Moffitt Cancer Center
Xiaoqing Yu, H. Lee Moffitt Cancer Center

Knowledge of the composition of a tumor can lead to more effective personalized treatment regimen. CIBERSORT employs linear support vector regression (SVR) to estimate the cell composition from gene expression profiles. Utilization of polynomial or radial kernels can lead to more accurate estimates than linear kernels. Tuning of these kernels is necessary for accurate estimates, but requires cross validation. Each model CIBERSORT constructs is unique to an individual; hence cross validation is not possible. Multiple kernel learning (MKL) can alleviate the need for cross validation and produce more accurate estimates by constructing a convex combination of kernel matrices. We present a novel approach to improve deconvolution techniques using MKL. We provide simulation results and apply it to reference gene expression profiles derived from single-cell RNA-seq data. By analyzing single-cell RNA-seq data of ~ 5000 cells from 20 head and neck tumors, we derived reference profiles for cell types associated with tumor microenvironment, including tumor, stromal, immune cells especially T cell sub-types.

email: cmwilson0109@gmail.com

INFORMATIVE PROJECTIONS AND DIMENSION REDUCTIONS FOR HIGH-DIMENSIONAL DATA CLUSTERING

Zhipeng Wang*, Genentech
David Scott, Rice University

Clustering is an unsupervised learning technique which groups data into "clusters" based on the similarity measure. Euclidean distance is the common similarity measure for a majority of clustering algorithms. But in the high-dimensional settings, the Euclidean distance becomes a poor similarity measure due to a large number of "noisy" features. Here we proposed a new algorithm for dimension reductions and variable selections to perform high-dimensional data clustering. We use projected distributions to determine the "informative features" for clustering, which are features whose projected distributions are mostly different from either uni-modal distribution or uniform distribution. We weighted all the features based on the KL divergence between the projected distributions and the referenced distributions, and randomly sample those features according to the weights to build a ensemble model. We performed the subspace clustering using the minimal spanning tree algorithm, spectral clustering, K-means, t-SNE and neural networks.

email: Zhipeng.Wang@alumni.rice.edu

SIMULTANEOUS ESTIMATION OF NUMBER OF CLUSTERS AND FEATURE SPARSITY IN CLUSTERING HIGH-DIMENSIONAL DATA USING RESAMPLING METHODS

Yujia Li*, University of Pittsburgh
Xiangrui Zeng, Carnegie Mellon University
Chien-Wei Lin, Medical College of Wisconsin
George Tseng, University of Pittsburgh

Estimating the number of clusters (K) is a critical and often difficult task in cluster analysis. Many methods have been proposed to estimate K, including some top performers using resampling approach. When performing cluster analysis in high-dimensional data, simultaneous clustering and feature selection is needed for improved interpretation and performance. To our knowledge, none has investigated simultaneous estimation of K and feature sparsity in an exploratory cluster analysis. In this paper, we propose a resampling method to meet this gap and demonstrate it under sparse K-means framework. Extensive simulations show its superior performance over classical method in estimating K in low-dimensional data. For high-dimensional data, our proposed method is also among the top performers in simultaneous estimation of K and feature sparsity in simulations. Finally, we apply the method to three leukemia transcriptomic applications. It achieves better clustering accuracy with fewer predictive genes, providing more biological insights to understand the identified disease subtypes.

email: yul178@pitt.edu

AN EVALUATION OF MACHINE LEARNING AND CLASSICAL STATISTICAL METHODS FOR DISCOVERY IN LARGE-SCALE TRANSLATIONAL DATA

Megan C. Hollister*, Vanderbilt University
Jeffrey D. Blume, Vanderbilt University

A recent paper in the top journal Nature Methods – "Statistics Versus Machine Learning" by Bzdok, Altman and Krzywinski – generated broad discussion and debate regarding the proper comparison of machine learning and traditional statistical methods in large-scale inference contexts. However, some of their methods were biased and led to the improper conclusion that random forests outperform traditional methods based on p-value adjustments. Here we reexamine those methods and provide a fair and unbiased comparison. We also take this opportunity to examine a new technique, second-generation p-values, and show it generally outperforms the other methods in question. The context is a simulated microarray of gene expression data for identifying dysregulated genes, which is representative of the large number of high dimensional analyses in the literature today. The results of our investigation shed light on how to choose methods to analyze large-scale translational data.

email: megan.c.hollister@vanderbilt.edu

37. SPACIO-TEMPORAL MODELING

SPATIO-TEMPORAL MODELS OF INTRAHEPATIC HEPATITIS C VIRUS PROPAGATION IN HUMANS

Paula Moraga*, Lancaster Medical School
Peter J. Diggle, Lancaster Medical School
Ruy M. Ribeiro, Universidade de Lisboa
Ashwin Balagopal, Johns Hopkins University
Benjamin M. Taylor, Lancaster Medical School

We develop spatio-temporal point process models to characterize the propagation patterns of Hepatitis C Virus in human livers. The models account for rates of hepatocyte infection, viral production, and immune responses. The sampling protocol generates individual cell locations for a patch of cells, together with measures of HCV RNA and expression levels of infection

inhibiting and facilitating genes. Within any one patch, we model the rate of infection from cell i to cell j at time t as a product of an innate force of infection, the infectivity of an infectious cell i, the susceptibility of an uninfected cell j, and the transmissitivity between cells i and j. In addition, infectivity and susceptibility are expressed as log-linear regressions on gene expression levels. To fit the model, we use a partial likelihood method which requires knowledge only of the ordering of the infection times, for which we substitute the ordering of the levels of HCV RNA. We calculate maximum partial likelihood estimates and compute 95% confidence intervals of the parameters using parametric bootstrap. The models developed help to understand the mechanics of infection in the liver.

email: p.e.moraga-serrano@lancaster.ac.uk

SIMULTANEOUS RANKING AND CLUSTERING OF SMALL AREAS BASED ON HEALTH OUTCOMES USING NONPARAMETRIC EMPIRICAL BAYES METHODS

Ronald E. Gangnon*, University of Wisconsin, Madison
Cora Allen-Coleman, University of Wisconsin, Madison

A common task is ranking different geographic units (small area), e.g. counties in the United States, based on health (or socioeconomic) outcomes/determinants. We propose a nonparametric empirical Bayes (finite mixture) model for small area health outcomes that is suitable for simultaneous ranking and clustering of small areas. Optimal point estimates of the ranks are obtained to minimize expected integrated squared error loss on the health outcome (mean or proportion) scale. Small areas are simultaneously clustered by assigning each small area to the modal cluster (mixture component) for its estimated rank position. We illustrate our approach using an analysis of percent low birth weight births for Wisconsin counties, 2008-2014.

email: ronald@biostat.wisc.edu

SPATIAL-TEMPORAL MODELS AND VALIDATION FOR PREDICTING HISTORICAL MISSING CANCER INCIDENCE

Benmei Liu*, National Cancer Institute, National Institutes of Health
Li Zhu, National Cancer Institute, National Institutes of Health
Huann-Sheng Chen, National Cancer Institute, National Institutes of Health
Joe Zou, IMS
Rebecca Siegel, American Cancer Society
Kim D. Miller, American Cancer Society
Ahmedin Jemal, American Cancer Society
Eric J. Feuer, National Cancer Institute, National Institutes of Health

Cancer registry data come from the NCI's Surveillance, Epidemiology and End Results (SEER) program and the CDC's National Program of Cancer Registries (NPCR). High-quality incidence data have not been achieved in all states historically due to data quality concerns and different releasing roles across different cancer registries. In addition, the total number of cases in the most recent 3 data years are incomplete because of delays in reporting. To predict the current year cancer incidence, historical missing incidence need to be imputed. This paper describes our research on applying spatial-temporal mixed effects models to impute the historically missing data. Different validation approaches of our modeled estimates will be also demonstrated.

email: benmeiliu@hotmail.com

USING SPATIOTEMPORAL MODELS TO GENERATE SYNTHETIC DATA FOR PUBLIC USE

Harrison Quick*, Drexel University
Lance Waller, Emory University

When agencies release public-use data, they must be cognizant of the potential risk of disclosure associated with making their data publicly available. This issue is particularly pertinent in disease mapping, where small counts pose both inferential challenges and potential disclosure risks. While the small area estimation, disease mapping, and statistical disclosure limitation literatures are individually robust, there have been few intersections between them. Here, we formally propose the use of spatiotemporal data analysis methods to generate synthetic data for public use. Specifically, we analyze ten years of county-level heart disease death counts for multiple age-groups using a Bayesian model that accounts for dependence spatially, temporally, and between age-groups; generating synthetic data from the resulting posterior predictive distribution will preserve these dependencies. After demonstrating the synthetic data's privacy-preserving features, we illustrate their utility by comparing estimates of urban/rural disparities from the synthetic data to those from data with small counts suppressed.

email: harryq@gmail.com

BAYESIAN SELECTION OF NEIGHBORHOOD STRUCTURE IN SPATIAL MODEL

Marie Denis*, CIRAD
Benoît Cochard, PalmElit

In the field of epidemiology, a common objective is to study the mapping of diseases in relation to space. The conditional autoregressive (CAR) model is the most popular approach allowing flexible modeling of the spatial dependence structure. These models allow considering spatial dependence by associating the outcome at a given site with those outcomes at neighboring sites according a neighborhood structure defined by the user. In many applications, the exploration of different neighborhood structures is of interest for gaining insights into the understanding of spread disease. In this objective we propose a Bayesian selection approach to select neighborhood structure in CAR model. A parametrization of neighborhood structure in the spirit of Zhu et al. (2010) combined with a Bayesian Shrinkage approach (Kyung et al., 2010) are used. We illustrate this approach on simulated datasets and with an application to the spread of infection in oil palm trees.

email: marie.denis@cirad.fr

38. EMERGING STATISTICAL ISSUES AND METHODS FOR INTEGRATING MULTI-DOMAIN MHEALTH DATA

METHODS FOR COMBINING ACTIVITY INFORMATION FROM HEART RATE AND ACCELEROMETRY IN THE BALTIMORE LONGITUDINAL STUDY OF AGING

Ciprian Crainiceanu*, Johns Hopkins University

Participants in the Baltimore Longitudinal Study of Aging completed a clinical assessment and wore an Actiheart monitor in the free-living environment. Actiheart

is a wearable device that measures both the heart rate and activity intensity continuously for weeks at a time. In this talk, I will discuss statistical methods for combining heart rate and accelerometry data to define periods of activity intensity and obtain meaningful summaries for downstream analyses. New analytical methods will be introduced for outcome prediction and analysis of daily patterns of activity using densely sampled multivariate predictors.

email: ccrainic@jhsph.edu

RAR: A REST-ACTIVITY RHYTHM DATA ANALYSIS PACKAGE IN R

Jessica Graves*, University of Pittsburgh
Haoyi Fu, University of Pittsburgh
Robert Krafty, University of Pittsburgh
Stephen Smagula, University of Pittsburgh
Matricia Hall, University of Pittsburgh

Features of circadian rhythms, or rest-activity rhythms (RARs), are associated with a variety of mental and physical health outcomes. Actigraphs are acceleration-based wearable devices that are increasingly utilized to characterize person-specific RAR features. Actigraphy's cost effectiveness and ability to noninvasively monitor subjects over long periods of time have led to its widespread use. Thus, it is important that analysis platforms are available to analyze these data. In this talk, we describe a new R package, "RAR", which analyzes person-specific RAR data. The aim of this package is to provide a standard, open-source platform for investigators to perform a comprehensive array of RAR analysis. These include classical parametric analysis, as well as recently developed semiparametric, localized, and `person-time' based analyses. The use of the package is illustrated by a study of depression in older adults in which RAR rhythmic variations and activity before typical rise times were associated with depression symptom severity. Our package will empower researchers to perform similar analyses easily, quickly, and with greater reproducibility.

email: jeg143@pitt.edu

MIXED EFFECTS NEURAL NETWORKS FOR LONGITUDINAL DATA

Ian J. Barnett*, University of Pennsylvania

While classical statistical models rely heavily on linear predictors, non-linear predictive deep learning models such as artificial neural networks have shown great potential in recent years. However, these methods tend to assume independent observations as input data and do not properly account for clustered or correlated observations. In order to account for clustered input data, we extend generalized lienar mixed models to include feed-forward neural networks in the fixed effects.

email: ibarnett@pennmedicine.upenn.edu

OBJECTIVE MEASUREMENT VERSUS PERFORMANCE TEST FOR PHYSICAL ACTIVITY: WHICH TO USE?

Jiawei Bai*, Johns Hopkins University
Vadim Zipunnikov, Johns Hopkins University
Lisa M. Reider, Johns Hopkins University
Daniel O. Scharfstein, Johns Hopkins University

Objective measurement of physical activity has become an important part of many studies that include assessment of human physical function. Accelerometer based wearable devices are the major tool utilized in such studies, because they can be worn on the human body relatively comfortably for an extended period of time – this enables a rich data collection in the free-living environment for weeks. However, some traditional methods such as performance tests are still used in many applications, because they are well-studied in the literature and often provide more detailed information on some specific function. In this paper, we introduce a series of statistical tools and models to assess and compare, on the same population, what information about the physical function we can get from free-living accelerometry measurement and in-lab performance tests. We used the data from the OUTLET Study of the METRC (Major Extremity Trauma Research Consortium), which aimed to compare 18-month functional outcomes and health related quality of life of patients undergoing salvage versus amputation following severe leg/foot injury.

email: jbai@jhsph.edu

39. CAUSAL INFERENCE WITH DIFFERENCE-IN-DIFFERENCES AND REGRESSION DISCONTINUITY DESIGNS

PATTERNS OF EFFECTS AND SENSITIVITY ANALYSIS FOR DIFFERENCES-IN-DIFFERENCES

Luke Keele*, University of Pennsylvania
Colin Fogarty, Massachusetts Institute of Technology
Jesse Hsu, University of Pennsylvania
Dylan Small, University of Pennsylvania

In the estimation of causal effects with observational data, applied analysts often use the differences-in-differences (DID) method. The method is widely used since the needed before and after comparison of a treated and control group is a common situation in the social and biomedical sciences. Researchers use this method since it protects against a specific form of unobserved confounding. Here, we develop a set of tools to allow analysts to better utilize the method of DID. First, we develop form of matching that allows for covariate adjustment under the DID identification strategy that is consistent with the hypothetical experiment. We also develop a well known method of sensitivity analysis for hidden confounding for the DID method. We develop these sensitivity analysis methods for both binary and continuous outcomes. We then apply our methods to two different empirical examples.

email: luke.keele@gmail.com

A BRACKETING RELATIONSHIP BETWEEN THE DIFFERENCE-IN-DIFFERENCES AND LAGGED DEPENDENT VARIABLE ADJUSTMENT

Fan Li*, Duke University
Peng Ding, University of California, Berkeley

Difference-in-differences is a widely-used evaluation strategy that draws causal inference from observational panel data. It crucially relies on the assumption of parallel trend, which is scale dependent and may be questionable in some applications. A common alternative method is a regression model that adjusts for the lagged dependent variable, which rests on the assumption of ignorability conditional on past outcomes. In the context of linear models, Angrist and Piskhe (2009) show that difference-in-differences and the lagged-dependent-variable regression analyses have a bracketing relationship. Namely, if ignorability is correct, then mistakenly assuming the parallel trend will overestimate a truly positive effect; in contrast, if the parallel trend is correct, then mistakenly assuming ignorability will underestimate a truly positive effect. This article proves that the same bracketing relationship holds in general nonparametric settings. We also extend the result to semiparametric estimation based on inverse probability weighting. We provide two real examples to demonstrate the theoretical result. This is a joint work with Peng Ding.

email: fl35@duke.edu

AUGMENTED WEIGHTING ESTIMATORS FOR DIFFERENCE-IN-DIFFERENCES

Frank Li*, Duke University
Fan Li, Duke University

Difference-in-differences (DID) is a widely used approach for drawing causal inference from observational panel data. Two common estimation strategies for DID are outcome regression and propensity score weighting. We focus on the common two-period two-group DID design and propose two augmented estimators that combine the virtue of regression and weighting. The first augmented DID estimator (DID1) requires an outcome model for both groups and is locally efficient when both the regression and propensity score models are correct. However, the consistency of DID1 strictly relies on a correct propensity score model. The second augmented DID estimator (DID2) specifies an outcome model only for the control group, and hence is not locally efficient. However, we show that DID2 is doubly robust. We further provide closed-form empirical sandwich variance estimators for these two DID estimators, and investigate their finite-sample performance via simulations. We apply the two DID estimators to the Pennsylvania Department of Transportation data and study the effectiveness of rumble strips in reducing vehicle crashes. This is joint work with Fan Li.

email: frank.li@duke.edu

A REGRESSION DISCONTINUITY APPROACH FOR ADDRESSING TEMPORAL CONFOUNDING IN THE EVALUATION OF THERAPEUTIC EQUIVALENCE OF BRAND AND GENERIC DRUGS

Ravi Varadhan*, Johns Hopkins University
Lamar Hunt, Johns Hopkins University
Dan Scharfstein, Johns Hopkins University
Irene Murimi, Johns Hopkins University
Jodi Segal, Johns Hopkins University

Generic drugs are required to be bioequivalent to their brand counterparts; yet, the generic approval process does not require demonstration of therapeutic equivalence. Methods are needed to assess therapeutic equivalence in situations where questions arise. A primary difficulty is that the dates of initiation for brand and generic users do not overlap. The use of brand drug plummets, when a generic enters the market. We develop a method to obtain causal estimates of the effectiveness of a generic compared to a brand product that accounts for temporal confounding in the presence of a positivity violation. Using venlafaxine, an antidepressant, we identified new users of brand and generic products within a Commercial Claims Data from 1994-2016. We apply regression discontinuity to survival curves with a discontinuity in the probability of initiation to generic at the date when generic becomes available. The survival curves are estimated using G-computation to adjust for time-varying confounding. Our method provides a comparison between the survival curves under adherent use of brand and generic, conditional on initiating treatment on the date of generic market entry.

email: ravi.varadhan@jhu.edu

40. STATISTICAL CHALLENGES IN SYNTHESIZING ELECTRONIC HEALTHCARE DATA

COMPARING REAL WORLD DATA WITH RANDOMIZED TRIAL RESULTS TO ASSESS VALIDITY: PRELIMINARY INSIGHTS FROM THE RCT DUPLICATE PROJECT

Jessica M. Franklin*, Brigham and Women's Hospital and Harvard Medical School
Sebastian Schneeweiss, Brigham and Women's Hospital and Harvard Medical School

Randomized controlled trials (RCTs) remain the gold standard for establishing the causal relationship between medications and health outcomes. However, for some clinical questions RCTs may be infeasible, unethical, costly, or generalizable to only a very narrow population. In these cases, observational studies from routinely collected "real-world" health data (RWD) are crucial for supplementing the evidence from RCTs, but concerns about the validity of real-world observational studies continue to detract from their utility for decision-making. To explore the validity of observational studies, we have launched RCT DUPLICATE, a large, comprehensive, prospective comparison of advanced observational RWD study approaches and RCTs, thereby providing guidance on how to optimize the performance of causal inference methods applied to RWD for the study of comparative effectiveness and safety of medications. In this talk, I will share the design and rationale of the project as well as initial learnings in RWD study implementation.

email: jmfranklin@bwh.harvard.edu

POPULATION-LEVEL EFFECT ESTIMATION: FROM ART TO SCIENCE

Martijn J. Schuemie*, Janssen R&D

When designing an observational study, there are many study designs to choose from, and many additional choices to make, and it is often unclear how these choices will affect the accuracy of the results. Here we present a new benchmark for evaluating population-level estimation methods. The benchmark consists of a gold standard of research hypothesis where the truth is known, and a set of metrics for characterizing a methods performance when applied to the gold standard. We distinguish between two types of tasks: (1) estimation of the average effect of an exposure on an outcome relative to no exposure (effect estimation), and (2) estimation of the average effect of an exposure on an outcome relative to another exposure (comparative effect estimation). The benchmark allows evaluation of a method on either or both tasks. We apply this benchmark to the OHDSI Methods Library, a set of R packages implementing most well-known observational analysis designs, such as the new-user cohort design and the self-controlled case series design. We evaluate a large number of variations of each design on a US insurance claims database.

email: schuemie@ohdsi.org

ROBUST PRIVACY-PRESERVING STATISTICAL METHODS FOR HORIZONTALLY PARTITIONED INCOMPLETE DATA IN DISTRIBUTED HEALTH DATA NETWORKS

Qi Long*, University of Pennsylvania
Changgee Chang, University of Pennsylvania
Yi Deng, Google
Xiaoqian Jiang, University of Texas Health Science Center at Houston

Distributed health data networks (DHDNs) such as pSCANNER that leverage electronic health records (EHRs) from multiple institutions have drawn substantial interests in recent years, as they eliminate the need to create, maintain, and secure access to central data repositories, mitigate many security, proprietary, and privacy concerns, and lower the hurdle to collaboration among multiple institutions. Missing data are ubiquitous in DHDNs, and need to be properly handled in statistical analysis. Leveraging recent development in distributed analysis methods, we develop robust privacy-preserving statistical methods for horizontally partitioned incomplete data in DHDNs that do not require pooling patient-level data into a centralized repository. Our simulation studies demonstrate that the proposed methods outperform several naïve methods and achieve similar performance as the existing missing data methods that use pooled data. Our methods are further illustrated using real data examples.

email: qlong@upenn.edu

OPPORTUNITIES AND CHALLENGES IN LEVERAGING REAL WORLD DATA IN REGULATORY CLINICAL STUDIES

Heng Li*, U.S. Food and Drug Administration
Lilly Q. Yue, U.S. Food and Drug Administration

In this era of "Big Data," there are many sources of real world healthcare data that could be leveraged in the clinical studies in the regulatory settings. While such large quantities of data reflect real world clinical practice and could potentially be used to reduce the cost or time of conducting clinical trials, statistical and regulatory challenges can emerge, particularly concerning the real-word data quality, the objectivity of investigational study design, and the reliability and interpretability of study results. This presentation will discuss such opportunities and challenges from statistical and regulatory perspectives.

email: heng.li@fda.hhs.gov

41. USING HISTORICAL DATA TO INFORM DECISIONS IN CLINICAL TRIALS: EVIDENCE BASED APPROACH IN DRUG DEVELOPMENT

LEVERAGING HISTORICAL CONTROLS USING MULTISOURCE ADAPTIVE DESIGN

Brian P. Hobbs*, Cleveland Clinic

Beneficial therapeutic strategies are established through a gradual process devised to define the safety and efficacy profiles of new strategies over a sequence of clinical trials. This system produces redundancies, whereby similar treatment strategies are replicated, either as experimental or comparator standard-of-care therapies, across development phases and multiple studies. This article describes a collection of web-based statistical tools hosted by MD Anderson Cancer Center that enable investigators to incorporate historical control data into analysis of randomized clinical trials using Bayesian hierarchical modeling as well as implement adaptive designs using the method described in Hobbs et al. (2013). By balancing posterior effective sample sizes among the study arms, the adaptive design attempts to maximize power on the basis of interim posterior estimates of bias. With balanced allocation guided by hierarchical modeling, the design offers the potential to assign more patients to experimental therapies and thereby enhance efficiency while limiting bias and controlling average type I error.

email: hobbsb@ccf.org

USE OF HISTORICAL DATA FOR PREMARKETING EVALUATION OF MEDICAL DEVICES

Nelson Lu*, U.S. Food and Drug Administration, Center for Devices and Radiological Health
Yunling Xu, U.S. Food and Drug Administration, Center for Devices and Radiological Health
Lilly Yue, U.S. Food and Drug Administration, Center for Devices and Radiological Health

Historical data, including prior clinical studies and real-world data such as devices or patient registries, are often used in supporting the premarket approval regulatory decisions for medical devices. In many of such applications, Bayesian hierarchical models and power prior approaches are naturally adopted to borrow information from the historical data in the study design and analysis. Sometimes, applicants conduct nonrandomized studies in which the results of a single warm study are compared with those of the control arm which is formed from historical data. To maintain the objectivity of study design and validity of study results, propensity score methodology is often utilized and proper procedure of study design need to be

implemented. This talk will present some examples and discuss some considerations and challenges from the statistical and regulatory perspectives.

email: nelson.lu@fda.hhs.gov

EVIDENCE SYNTHESIS OF TIME-TO-EVENT DATA IN DESIGN AND ANALYSIS OF CLINICAL TRIALS

Satrajit Roychoudhury*, Pfizer Inc.

Bayesian methods have generated extensive interest in clinical trial design and analysis. Recently the 21st Century Cures Act and PUDUFA VI encourage the use of Bayesian techniques to accelerate drug development. Enriching control arm of new trial with relevant trial external information holds the promise of more efficient trial design in many occasions. Use of trial external information allows effective trial design including smaller sample size. One appeal of Bayesian approach is incorporation of historical data into the statistical analysis as "prior". However, the use of all available trial-external relevant information into prior is challenging as the information may accumulate in parallel to current trial. We propose an approach to use relevant trial external control data for designing trials with time-to-event endpoints. The proposed approach uses meta-analytic framework by considering the between-trial variability. It is flexible to incorporate individual as well as aggregated co-data and differential discounting. We use a phase II Oncology trial design to illustrate the methodology along with essential practical aspects.

email: satrajit.roychoudhury@pfizer.com

42. STATISTICAL INNOVATIONS IN SINGLE-CELL GENOMICS

CHARACTERIZING TECHNICAL ARTIFACTS IN SINGLE-CELL RNA-Seq DATA USING A DATA GENERATION SIMULATION FRAMEWORK

Rhonda Bacher*, University of Florida
Christina Kendziorski, University of Wisconsin, Madison
Li-Fang Chu, Morgridge Institute for Research
Ron Stewart, Morgridge Institute for Research

Single cell RNA-sequencing (scRNA-seq) is a promising tool that facilitates study of the transcriptome at the resolution of a single cell. However, along with the many advantages of scRNA-seq come technical artifacts not observed in bulk RNA-seq studies including an abundance of zeros, varying levels of technical bias across gene groups, and systematic variation in the relationship between sequencing depth and gene expression. We previously developed the normalization method SCnorm to account for variability in the relationship between expression and sequencing depth, which we refer to as the count-depth relationship. To investigate the source of this variability, we developed a first principles simulation framework which takes each step of generating scRNA-seq data into account. With this framework, we demonstrate the contribution of various protocol choices to technical artifacts observed in scRNA-seq data. Furthermore, we illustrate how a critical step in most scRNA-seq protocols directly contributes to the systematic variability in the count depth relationship, and show that hypotheses generated with the simulation are supported by existing independent datasets.

email: rbacher@ufl.edu

HI-C DECONVOLUTION VIA JOINT MODELING OF BULK AND SINGLE-CELL HI-C DATA

Yun Li*, University of North Carolina, Chapel Hill
Ruth Huh, University of North Carolina, Chapel Hill
Yuchen Yang, University of North Carolina, Chapel Hill
Jin Szatkiewicz, University of North Carolina, Chapel Hill

Hi-C allows genome-wide study of the genome's 3D structure. Standard Hi-C data derive from millions of cells, providing a population average measure of heterogeneous cells. Many deconvolution methods exist for bulk RNA-seq data. However, there are no deconvolution methods for Hi-C data. Here, we adopt a matrix decomposition framework to estimate cell type proportions in bulk Hi-C samples. While in RNA-seq data, gene expression levels are used to profile the cells, different levels of Hi-C readouts such as metrics for topologically associating domains, interchromosomal and intrachromosomal contacts, exist. We have performed extensive real data based simulations to assess our method, including mixing HAP1 and HeLa (two human cancer cell lines) single cells, at varying (10-90%) proportions; and mixing mix two mouse cell lines: Patski and MEF. Using interchromosomal contacts, our method provides accurate estimation of the mixture. Absolute deviations of estimated proportions range from 0.001 to 0.09 for the human dataset and 0.024 to 0.092 for the mouse dataset. We aim to assess our method using different Hi-C metrics to provide the optimal solution for cell type proportions.

email: yunli@med.unc.edu

SEMI-SOFT CLUSTERING OF SINGLE CELL DATA

Kathryn Roeder*, Carnegie Mellon University
Lingxue Zhu, Carnegie Mellon University
Bernie Devlin, University of Pittsburgh
Jing Lei, Carnegie Mellon University
Lambertus Klei, University of Pittsburgh

Motivated by the dynamics of development, in which cells of recognizable types, or pure cell types, transition into other types over time, we propose a method of semi-soft clustering that can classify both pure and intermediate cell types from data on gene expression from individual cells. Called SOUP, for Semi-sOft clUstering with Pure cells, this novel algorithm reveals the clustering structure for both pure cells and transitional cells with soft memberships. SOUP involves a two-step process: identify the set of pure cells and then estimate a membership matrix. To find pure cells, SOUP uses the special block structure in the expression similarity matrix. Once pure cells are identified, they provide the key information from which the membership matrix can be computed. By modeling cells as a continuous mixture of K discrete types we obtain more parsimonious results than obtained with standard clustering algorithms. Moreover, using soft membership estimates of cell type cluster centers leads to better estimates of developmental trajectories. The strong performance of SOUP is documented via simulation studies and analyses of two data sets.

email: roeder@stat.cmu.edu

EVALUATION OF CELL CLUSTERING IN SINGLE CELL DATA

Zhijin (Jean) Wu*, Brown University

Unsupervised clustering is a common approach to discovering latent structure in high dimensional data. It has been successfully applied to genomic data in identifying, for example, subtypes of cancer and new cell types. It is widely used in single cell genomics data in visualization as well as in discovering new cell types or subtypes. A challenge in the choice of clustering methods is the difficulty of evaluating the performance of unsupervised clustering methods. This challenge becomes more serious in cell type clustering, even when there is some knowledge of 'true clusters", as existing metrics often treat the clusters as exchangeable and fail to recognize the natural hierarchy in cell type and subtypes. In this presentation we discuss the interpretation, the pros and cons of common metrics of clustering performance in the application of single cell data, and present novel measures designed for cell clustering.

email: zhijin_wu@brown.edu

43. ADVANCES IN STATISTICAL METHODS FOR SURVEILLANCE DATA OF INFECTIOUS DISEASES

STATISTICAL CHALLENGES WHEN ANALYSING EMERGING EPIDEMIC OUTBREAKS

Tom Britton*, Stockholm University
Gianpaolo Scalia Tomba, University of Rome Tor Vergata

New infectious disease outbreaks have great impact on communities over the world, as recently manifested by the Ebola outbreak. An important statistical task is then to predict the future scenario with and with out preventive measures. In the current talk we will investigate such analyses and see how it can be improved. The main catch is that in the exponentially growing phase early on in an outbreak, several biases can occur if not taken account for: events with short delays will be over-represented. We will give some examples from the Ebola outbreak and see how the biases can be removed or at least reduced. (Joint work with Gianpaolo Scalia Tomba).

email: tom.britton@math.su.se

EFFICIENT BAYESIAN SEMIPARAMETRIC MODELING AND VARIABLE SELECTION FOR SPATIO-TEMPORAL TRANSMISSION OF MULTIPLE PATHOGENS

Nikolay Bliznyuk*, University of Florida
Xueying Tang, Columbia University

An epidemic of an infectious disease is often a consequence of the circulation of multiple pathogens. Modeling pathogen-specific transmission is helpful to the design of public health policy and intervention strategies. Since pathogen information is often not available from national surveillance, a small subset of cases undergoes laboratory testing to identify the pathogen. Appropriate imputation of the missing pathogen information for the majority of cases is usually very computationally intensive, if feasible at all. We propose an efficient hierarchical Bayesian model to characterize epidemics of multiple pathogens. Here, imputation of the unknown pathogen-specific cases can be avoided by exploiting the relationship between multinomial and Poisson distributions. A variable selection prior is also employed to identify the risk factors and their proper functional form respecting the linear-nonlinear hierarchy. The efficiency gains of the proposed model and the performance of the selection priors are investigated through simulation studies. Our model is applied to a subset of the hand, foot and mouth disease data from 2009 in mainland China.

email: nbliznyuk@ufl.edu

DYNAMIC MONITORING OF SPATIO-TEMPORAL DISEASE INCIDENCE RATES

Peihua Qiu*, University of Florida

Online sequential monitoring of the incidence rates of infectious diseases is critically important for public health. Governments around the world have invested a great amount of resource in building global, national and regional disease reporting and surveillance systems. In these systems, conventional control charts, such as the CUSUM and EWMA charts, are routinely included for disease surveillance purpose. However, these charts require many assumptions on the observed data, including the ones that the observed data are independent and identically normally distributed when no disease outbreaks are present. These assumptions are rarely valid in practice, making the results from the conventional control charts unreliable. We develop a new sequential monitoring approach which can accommodate the dynamic nature of the disease incidence rates (i.e., seasonality), spatio-temporal data correlation, and non-normality. It is shown that the new method is much more reliable to use in practice than the commonly used conventional charts for sequential monitoring of disease incidence rates.

email: pqiu@ufl.edu

STATISTICAL ADJUSTMENT FOR REPORTING BIAS IN SURVEILLANCE DATA OF INFECTIOUS DISEASES

Yang Yang*, University of Florida
Tim Tsang, University of Florida
Diana Rojas Alvarez, University of Florida
Ira Longini, University of Florida
M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center

Reporting bias is common in the surveillance of infectious diseases, often a result of efforts to identify the most vulnerable subpopulation. We investigated reporting bias at two levels. The institutional level is motivated by the Chinese influenza outbreak investigation data, where outbreaks in institutes (e.g., schools) were reported only if there were 30 or more influenza cases. For proper inference, the likelihood for transmission events need to be conditioned on the fact that the observed final size of the outbreak exceeds the lower limit of detection. However, the distribution of final size becomes numerically challenging to calculate even for moderate sizes of clusters. At the population level, the Zika surveillance data from Colombia is a typical example, where female cases in reproductive age were much more actively sought than other groups because of the link between Zika infection and microcephaly. We are interested in estimating the bias-corrected basic reproductive number for Zika. We discuss newly developed methods for addressing these challenges, simulation results to validate these methods, and their application to real surveillance data.

email: yangyang@ufl.edu

44. SPEED POSTERS: SPATIO-TEMPORAL MODELING/ LONGITUDINAL DATA/SURVIVAL ANALYSIS

44a. INVITED SPEED POSTER: A MULTIVARIATE SPATIO-TEMPORAL MODEL OF OPIOID OVERDOSE DEATHS IN OHIO

Staci Hepler*, Wake Forest University
David Kline, The Ohio State University

In 2015, Ohio led the nation in fatal overdoses due to opioid misuse. In earlier years it was believed prescription opiates were driving the opioid crisis in Ohio. However, following policy changes, opioid overdose deaths due to heroin and fentanyl have drastically increased. In this work, we develop a Bayesian multivariate spatio-temporal model for Ohio county overdose death rates from 2007 to 2016 due to different types of opiates. We jointly model death rates due to multiple types of opiates using a spatial rates model with a latent multivariate conditional autoregressive process. We incorporate change point regression to identify large-scale shifts within each type of opiate considered. This model allows us to not only study how socio-environmental factors relate to opioid overdose deaths, but also spatio-temporal trends in the types of opioids contributing to death.

email: heplersa@wfu.edu

44b. INVITED SPEED POSTER: A MULTIVARIATE DYNAMIC SPATIAL FACTOR MODEL FOR SPECIATED POLLUTANTS AND ADVERSE BIRTH OUTCOMES

Kimberly A. Kaufeld*, Los Alamos National Laboratory

Evidence suggests that exposure to elevated concentrations of air pollution during pregnancy is associated with increased risks of birth defects and other adverse birth outcomes. While current regulations put limits on total PM2.5 concentrations, there are many speciated pollutants within this size class that likely have distinct effects on perinatal health. However, due to correlations between these speciated pollutants, it can be difficult to decipher their effects in a model for birth outcomes. To combat this difficulty, we develop a multivariate spatio-temporal Bayesian model for speciated particulate matter using dynamic spatial factors. These spatial factors can then be interpolated to the pregnant women's homes to be used to model birth defects. The birth defect model allows the impact of pollutants to vary across different weeks of the pregnancy in order to identify susceptible periods. The proposed methodology is illustrated using pollutant monitoring data from the Environmental Protection Agency and birth records from the National Birth Defect Prevention Study.

email: kkaufeld@lanl.gov

44c. FORECASTING CANCER INCIDENCE AND MORTALITY UNDER THE AGE-PERIOD-COHORT MODEL

Ana F. Best*, National Cancer Institute, National Institutes of Health
Philip S. Rosenberg, National Cancer Institute, National Institutes of Health

Age-Period-Cohort (APC) models are widely used for the study of disease incidence rates, particularly cancer surveillance research. The APC model characterizes rates into age effects reflecting the natural history of the disease of interest, period effects impacting all ages simultaneously, and cohort effects representing risk differences across birth years. Recent papers have used APC methods to forecast cancer incidence rates, and cause-specific death rates. We introduce methods to simultaneously forecast cancer incidence and mortality rates and counts using the NCI's Surveillance, Epidemiology, and End Results (SEER) program database, which tracks both cancer incidence and all-cause mortality of cancer cases. We forecast incidence by extrapolating a piecewise log-linear spline of the cohort rate-ratio, and the global curvature for period (or period rate-ratio and global curvature for cohort). For mortality, we additionally extrapolate a piecewise linear spline for the survival effect of diagnosis year. We assess the use of model averaging to incorporate model selection uncertainty into prediction intervals, and illustrate our methods using SEER data on Multiple Myeloma.

email: ana.best@nih.gov

44d. A HIERARCHICAL BAYESIAN APPROACH TO PREDICTING TIME-TO-CONVERSION TO ALZHEIMER'S DISEASE USING A LONGITUDINAL MAP OF CORTICAL THICKNESS

Ning Dai*, University of Minnesota
Hakmook Kang, Vanderbilt University
Galin Jones, University of Minnesota
Mark Fiecas, University of Minnesota

Prior studies have shown that cortical atrophy is associated with an increased risk of progression to dementia. In this work, we use the longitudinal structural magnetic resonance imaging (MRI) data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to investigate the relationship between conversion from mild cognitive impairment (MCI) to Alzheimer's disease (AD), and the dynamically changing cortical thickness over time and across the cortical surface. We develop a novel hierarchical Bayesian framework that embeds a spatial model on the cortical thickness effects within a survival model that predicts the time to progress to AD. The proposed method allows for interpretation with respect to the temporal dynamics of imaging measurements, identifies the patterns of regions where atrophy is associated with AD, and improves the performance of prediction by exploiting the spatial structure underlying the high-resolution observations over the cortical surface. We apply the proposed method to the longitudinal MRI data from ADNI to investigate how the impact of cortical thinning on the time to progress to AD varies across different regions of the brain.

email: daixx224@umn.edu

44e. BODY MASS INDEX AND BREAST CANCER SURVIVAL: A CENSORED QUANTILE REGRESSION ANALYSIS

Dawen Sui*, University of Texas MD Anderson Cancer Center
Mary Elizabeth Edgerton, University of Texas MD Anderson Cancer Center

BMI has been an important factor affecting breast cancer outcomes. Traditionally the associations of survival are investigated using Cox regression (CR). This may result in biases because of the assumption of CR. In this study, a distribution-free method, Censored Quantile Regression (CQR), is employed to evaluate the associations of survival with BMI and the results are compared to CR. Charts of 3008 patients at MD Anderson diagnosed from 1995 to 2005 were reviewed. BMI was categorized

as underweight (10%); normal weight (34%); overweight (28%); and obese (28%). BMI was significantly associated with overall survival (P<0.01) but not with disease relapse-free survival. The CQR and Cox models had consistent results showing that patients with normal weight had the lowest mortality risk, followed by overweight, obese, and underweight respectively. Multivariate analyses generated similar conclusions. CQR provides an alternative, more flexible way to analyze survival data and shows the detailed impact response during dynamic survival time. The Cox and CQR models together can provide more comprehensive information about breast cancer survival rate.

email: dawensui@mdanderson.org

44f. SUBGROUP ANALYSES USING COX REGRESSION: RELAXING THE PROPORTIONAL HAZARDS (PH) ASSUMPTION

Karen Chiswell*, Duke Clinical Research Institute
Amanda Brucker, Duke Clinical Research Institute
Adrian Coles, Duke Clinical Research Institute
Yuliya Lokhnygina, Duke University
Megan L. Neely, Duke University
Adam Silverstein, Duke Clinical Research Institute
Daniel M. Wojdyla, Duke Clinical Research Institute

Advice about analyzing heterogeneity of treatment effects in clinical trials is unanimous: subgroup analyses should be based on statistical tests for interaction between treatment and subgroup levels, the results of which should be reported along with effect estimates including confidence intervals (e.g., Wang et al, NEJM, 2007). For time-to-event endpoints, analyses are often conducted using the Cox proportional hazards (PH) regression model. We will consider two models for including subgroup effects and interactions. Model 1 (the standard approach) assumes that the hazard functions for the groups defined by the combinations of treatment x subgroup are all proportional to one another. Model 2 (stratified Cox regression) assumes PH for treatment groups within each subgroup, but allows for non-PH across subgroups. Using simulation we will evaluate the performance of Model 1 vs. Model 2 for analyzing data generated under various scenarios of non-PH. Impact of model assumptions on the size and power of interaction tests, and bias of subgroup-specific hazard ratio estimates will be summarized.

email: karenc2204@yahoo.com

44g. COMPARISON OF METHODS FOR ANALYZING TIME-VARYING CONTINUOUS COVARIATES IN SURVIVAL ANALYSIS

Qian Liu*, Arbor Research Collaborative for Health
Abigail R. Smith, Arbor Research Collaborative for Health
Laura H. Mariani, University of Michigan
Viji Nair, University of Michigan
Jarcy Zee, Arbor Research Collaborative for Health

Longitudinal biomarkers are useful for tracking chronic disease progression, but their values are known only at the exact time of measurement. We compared four methods that have been applied in clinical studies to analyze time-varying continuous covariates in Cox models: 1) time-invariant baseline, that only uses the single baseline value, 2) time-invariant average, that averages values over all follow-up and uses the average as a baseline covariate, 3) time-varying last observation carried forward (LOCF), that assumes the covariate is unchanged until the next observed value, and 4) time-varying cumulative average, that updates the average using values at or before each measurement. Simulations were used to compare methods under different true effect mechanisms and different covariate trajectories. The time-invariant average method had large bias and time-invariant baseline method was slightly biased toward the null. Time-varying LOCF and cumulative average methods were slightly biased when the other was the true mechanism, with the direction of bias driven by both measurement error and scaling effects. Finally, we compared these methods with data from the NEPTUNE study.

email: Jarcy.Zee@arborresearch.org

44h. JOINT MODELLING OF SEQUENTIAL TIME-TO-EVENTS

Akhtar Hossain*, University of South Carolina, Columbia
Hrishikesh Chakraborty, Duke Clinical Research Institute

In many clinical studies, dependent time-to-events data may be observed in a fashion where a specific event can only be observed if and only if the subject has experienced a previous event. Such events can be referred to as sequential time-to-events as they occur in a certain sequence. Analysis of sequential time-to-events can be very beneficial in predicting complex disease processes as well as understanding the correlation between survival events in question. There are vast of studies in literature focusing analysis of correlated time-to-events. However, literature seriously lacks a well-suited method for analyzing sequential time-to-events. This paper introduces a joint modeling framework to analyze sequential time-to-events and longitudinally observed outcomes. The estimation and significance test of parameters are discussed. Simulation studies show that the proposed method provides unbiased and reliable estimates of parameters. An application of the proposed model is demonstrated to model the transition times between undetectable and detectable viral loads using South Carolina (SC) HIV/AIDS surveillance data.

email: mhossain@email.sc.edu

44i. AN ALTERNATIVE TO THE COX MODEL ESTIMATOR IN TIME-TO-EVENT CLINICAL TRIALS WITH TREATMENT OF PHYSICIAN'S CHOICE

Philani B. Mpofu*, Indiana University, Richard M. Fairbanks School of Public Health
Stella W. Karuri, U.S. Food and Drug Administration

Clinical trials with treatment of physician's choice (TPC) are increasingly being used to study the efficacy of oncology treatments. In this design, the study treatment of interest is compared to a control group that consists of multiple treatments that are prescribed to participating patients at the discretion of their physicians. For purposes of statistical analysis, all the control treatments are often pooled into a single entity, and the Cox model based hazard ratio is used to estimate the treatment effect. We show that the pooling of treatments in the control into a single entity results in the violation of the proportional hazards (PH) assumption, except when all the control arm treatments are equally effective relative to the study treatment. Due to the violation of the PH assumption, the hazard ratio estimated through the usual Cox model is an average hazard ratio over the study time. Moreover, this average hazard ratio estimate is dependent on the censoring distribution. We propose an alternative

estimator for the average hazard ratio that is independent of censoring. We study the finite sample properties of the alternative estimator using simulations.

email: phmpofu@iupui.edu

44j. PHASE II TRIAL DESIGN WITH GROWTH MODULATION INDEX AS THE PRIMARY ENDPOINT

Jianrong John Wu*, University of Kentucky

Molecularly targeted, genomic-driven and immunotherapy-based clinical trials continue to be advanced for the treatment of relapse or refractory cancer patients, where the growth modulation index (GMI) is often considered a primary endpoint of treatment efficacy. However, there little literature is available that consider the trial design with GMI as the primary endpoint. In this article, we derived a sample size formula for the score test under a log-linear model of the GMI. Study designs using the derived sample size formula are illustrated under a bivariate exponential model, the Weibull frailty model and the generalized treatment effect size. The proposed designs provide sound statistical methods for a single-arm phase II trial with GMI as the primary endpoint.

email: jianrong.wu@uky.edu

44k. MULTILEVEL STOCHASTIC BLOCKMODEL FOR DYNAMIC NETWORKS

Jihui Lee*, Weill Cornell Medicine
Jeff Goldsmith, Columbia University
Gen Li, Columbia University

Analyzing dynamic networks provides an important perspective in understanding the topological evolution of relational data over time. In this study, we propose a method that directly parametrizes the connectivity between nodes as a smooth function of time. The proposed method is especially flexible in analyzing time-varying networks and its smooth parameters can be interpreted as evolving strength of engagement. When multiple networks are densely observed, they are often summarized and simplified as a single network. The proposed method instead accommodates the multiple observations within a time frame as a form of random effects. Essentially, it is a generalized version of stochastic blockmodels with a priori block membership with temporal random effects. It allows a variability among multiple relational patterns and provides a richer representation of dependent engagement patterns at each time point.

email: jil2043@med.cornell.edu

44l. INTEGRATIVE VARIABLE SELECTION METHOD FOR SUBGROUP ANALYSES IN LONGITUDINAL DATA

Xiaochen Li*, Indiana University
Sujuan Gao, Indiana University

Longitudinal data are often collected in modern medical studies. With the improvements of technology, researchers are able to collect information on an increasing number of predictors which presents the statistical challenge of variable selection. We propose a three-stage, model-based method to select informative factors and two-way interactions which is crucial for subgroup identification in longitudinal clinical trials. At the first stage, we use marginal score tests to select variables associated with the longitudinal outcome. At the second stage, we use covariance-insured screening to identify variables associated with those selected during the first step. In the third stage, we apply a penalized LASSO method using the variables in steps 1 and 2 to obtain all informative variables and their interactions. Simulation studies are conducted to evaluate the performance of the proposed method. A longitudinal clinical trial study is used to illustrate the proposed method.

email: li298@iu.edu

44m. EVALUATION OF THE PERFORMANCE OF PROPENSITY SCORE WEIGHTING METHODS FOR SURVIVAL OUTCOMES

Uma Siangphoe*, Janssen Research & Development
Kwan R. Lee, Janssen Research & Development

Propensity score weighting methods have been increasingly used in observational and real-world evidence studies to reduce potential confounding bias that could be introduced by imbalance of covariates between treatment exposures particularly for survival outcomes. The performance of the different weighting methods can depend on data features, propensity score distributions, and survival distributions. This study evaluated, through simulation, different propensity score weights with their different applications (e.g., truncated weighting methods and variance estimation of causal effect). We created simulation scenarios with different propensity score and survival distributions at various risk levels. We assessed the performance of the different weighting methods based on bias, mean squared error, and coverage probability (95% confidence interval) of the estimands. In addition to simulation study, an application to real-world example has been included for demonstration and illustration.

email: uma.siangphoe@gmail.com

45. BIOMARKERS

GROUP TESTING ESTIMATION FOR MULTIPLE INFECTIONS WITH ADJUSTMENTS FOR DILUTION EFFECTS

Md S. Warasi*, Radford University
Hrishikesh Chakraborty, Duke University

Group testing, which involves pooling individual biospecimens, is commonly used as a cost-effective testing method in epidemiological applications, for both disease screening and disease prevalence estimation. In this article, we study the problem of estimating the prevalence of multiple correlated diseases when there is a potential risk of model misspecification due to pooled dilution effects. To accomplish the estimation while adjusting for the adverse effect of dilution, we model the disease-specific assay sensitivity and specificity by continuous biomarker distributions in addition to modeling the binary pooled responses. Estimation is then achieved by the maximum likelihood technique. We demonstrate through simulation and real data

applications that our joint modeling approach offers reliable estimates for the disease prevalence rates whether a dilution effect is present or not. For dissemination purposes, we develop user-friendly R codes.

email: msarker@radford.edu

NONPARAMETRIC CONDITIONAL DENSITY ESTIMATION FOR POOLED BIOMARKER DATA

Dewei Wang*, University of South Carolina
Xichen Mou, University of South Carolina
Joshua Tebbs, University of South Carolina

In biomarker studies, when resources (e.g., the budget and/or the number of specimens) are limited, pooling specimens to take measurements of the biomarker's concentration level is often an alternative means. This article develops a kernel-based regression estimator of a biomarker level's density when a continuous covariate is available for each specimen but the biomarker is measured in pools with measurement errors. Consistency and asymptotic normality of our estimator is established. The rates of convergence depend on the tail behavior of the characteristic functions of the measurement error and the biomarker level. Simulation studies demonstrate the practical advantages of our method when comparing to the existing work. We further illustrate our method via a Polyfluorochemical data set.

email: deweiwang@stat.sc.edu

A MIXTURE OF TUKEY'S g-h DISTRIBUTIONS WITH APPLICATION TO MEASURING PROTEIN BIOMARKERS

Tingting Zhan*, Thomas Jefferson University
Inna Chervoneva, Thomas Jefferson University

Tukey g-h distributional family, generated from a simple transformation of the standard normal distribution, is well-known to approximate skewed and heavy tailed unimodal distributions, but not suitable for bimodal distributions. We introduce a two-component Tukey g-h mixture to accommodate bi-modality, skewness and heavy tails. The mixture parameters are estimated using an indirect estimator minimizing the distance between sample and model quantiles, thus avoiding the complexity and computational burden associated with approximating the likelihood function. The resulting estimator is fast convergent and robust to potential outliers. The proposed two-component Tukey g-h mixture was used to model the distributions of quantitative immunofluorescence immunohistochemistry (QIF-IHC) protein expressions in cancer cells of breast cancer tumor tissues. The mean of the second component was used as a potential biomarker, while the first component was considered representing the background immunofluorescence levels. This approach identified novel biomarkers that do not have a prognostic value when the standard average mean signal intensity is considered.

email: tingting.zhan@jefferson.edu

BAYESIAN NONPARAMETRIC CLUSTERING ANALYSIS FOR MULTI-SCALE MOLECULAR DATA

Yize Zhao*, Weill Cornell Medicine

Investigating cancer genome based on multi-type omics data is a global medical issue. Though some of the existing clustering methods are capable to character certain degree of concordant and heterogeneity across data types, none of them has incorporated biological network information within and across molecular modalities under cancer subtype discovery. Meanwhile, it is biologically important to identify the core set of biomarkers that are informative to the similarity among samples in each subtype. In this work, we construct a unified clustering model with an incorporation of biological network within and across different molecular data types and simultaneously identifying informative molecular biomarkers for each subtype. Different from existing parametric methods, we adopt a nonparametric approach based on Bayesian Dirichlet process mixture (DPM) models, which is more adaptable to different data types, robust to statistical assumptions and has no constrain on the number of clusters. The performance of the proposed model has been assessed by extensive simulation studies and read data application based on The Cancer Genome Atlas (TCGA) Research Network.

email: yiz2013@med.cornell.edu

46. BAYESIAN MODELING AND VARIABLE SELECTION

BAYESIAN SPARSE ENVELOPE MODEL FOR MULTIVARIATE LINEAR REGRESSION

Minji Lee*, University of Florida
Saptarshi Chakraborty, Memorial Sloan Kettering Cancer Center
Zhihua Su, University of Florida
Malay Ghosh, University of Florida

We propose the Bayesian sparse envelope model in the context of multivariate linear regression. The Bayesian sparse envelope model can conduct variable selection on the responses and achieve the efficiency gains compared to the standard Bayesian model. We demonstrate that our method performs well through simulations and data analysis. We obtain the asymptotic normality of the posterior distribution.

email: mlee9@ufl.edu

INTERACTION DETECTION USING BAYESIAN DECISION TREE ENSEMBLES

Junliang Du*, Florida State University
Antonio R. Linero, Florida State University

Methods based on Bayesian decision tree ensembles have proven valuable in constructing high-quality predictions, and are particularly attractive in certain settings because they encourage low-order interaction effects. Despite adapting to the presence of low-order interactions for prediction purpose, we show that Bayesian decision tree ensembles are generally anti-conservative for the purpose of conducting interaction detection. We address this problem by introducing Dirichlet process forests (DP-Forests), which leverage the presence of low-order interactions by clustering the trees so that trees within the same cluster focus on detecting a specific interaction. We

show on both simulated and benchmark data that DP-Forests perform well relative to existing interaction detection techniques for detecting low-order interactions, attaining very low false-positive and false-negative rates while maintaining the same performance for prediction using a comparable computational budget.

email: moonly.jp@gmail.com

BAYESIAN SELECTION OF VARIANCE COMPONENTS IN LINEAR MIXED MODELS

Benjamin Heuclin*, Université de Montpellier, France
Marie Denis, CIRAD, France
Frédéric Mortier, CIRAD, France
Catherine Trottier, Université de Montpellier, France

Linear mixed models are flexible tools for modeling a wide range of data types in various applied fields such as medicine, genetic or ecology. However, a key aspect in statistical analysis is model selection. While many approaches have already been proposed for the selection of the fixed effects part, only few methods are available for the identification of the non-zero variance components. This is much more difficult due to boundary problems arising from positive semi-definite constraints on covariance matrices. For longitudinal data, some works in frequentist and Bayesian context have been developed (Chen and Dunson (2003)). Recently, Lu et al. (2015) developed a Spike and Slab prior for standard deviations in variance component models. In their work, the slab prior distribution allows negative values.In this talk, we propose alternative prior distributions ensuring non-negative values for the standard deviations in the spirit of Bayesian shrinkage approaches and Spike and Slab methods. MCMC algorithms are developed to infer parameters. These new model formulations are compared using simulated data and applied to real dataset in the context of genetic association studies.

email: benjamin.heuclin@umontpellier.fr

BAYESIAN HIERARCHICAL MODELING ON COVARIANCE VALUED DATA

Satwik Acharyya*, Texas A&M University
Zhengwu Zhang, University of Rochester
Anirban Bhattacharya, Texas A&M University
Debdeep Pati, Texas A&M University

Analysis of functional connectivity of human brains is of pivotal importance for diagnosis of cognitive ability. The Human Connectome Project(HCP) provides neural data across different regions of interest of the living human brain. Individual specific data were available from an analysis(Dai et al., 2015) in the form of time varying covariance matrices representing the brain activity as the subjects perform a specific task. As a preliminary objective of studying the heterogeneity of brain connectomics across the population, we develop a probabilistic model for a sample of covariance matrices using a scaled Wishart distribution. Based on empirical explorations suggesting the data matrices to have low effective rank, we further model the center of the Wishart distribution using an orthogonal factor model type decomposition. We encourage shrinkage towards a low rank structure and discuss strategies to sample from the posterior distribution using a combination of Gibbs and slice sampling. We extend our modeling framework to a dynamic setting to detect change points and efficacy rates are checked on several case studies including our motivating HCP data.

email: satwik91@gmail.com

KNOWLEDGE-GUIDED BAYESIAN VARIABLE SELECTION IN NON-LINEAR SUPPORT VECTOR MACHINES FOR STRUCTURED HIGH-DIMENSIONAL DATA

Wenli Sun*, University of Pennsylvania
Changgee Chang, University of Pennsylvania
Qi Long, University of Pennsylvania

Support vector machines (SVM) is a popular classification method for analysis of high dimensional data such as genomics data. Recently, many linear SVM methods with feature selection have been developed under frequentist regularization or Bayesian shrinkage. However, the linear dependency assumption may not be realistic but few works apply feature selection in nonlinear Bayesian SVM models. By using the ideas of Gaussian processes, Henao et al develop a nonlinear kernelized Bayesian SVM to bridge the gap between linear and nonlinear Bayesian SVM models. Building on this framework, we apply Ising priors into the kernel and enable feature selection guided by the knowledge on the graphical structure among predictors, e.g, biological pathways among genes. The performance of our method is evaluated and compared with existing linear and nonlinear SVM methods in terms of prediction and feature selection in extensive simulation settings. In addition, our method is illustrated in the analysis of genomic data from a cancer study, demonstrating its advantage in generating biologically meaningful results and identifying potentially important features.

email: wenlisun@pennmedicine.upenn.edu

BAYESIAN ADJUSTMENT FOR CONFOUNDING WHEN ESTIMATING AVERAGE CAUSAL EFFECTS FOR TIME-TO-EVENT OUTCOMES

Li Xu*, University of Kentucky
Chi Wang, University of Kentucky

The Bayesian adjustment for confounding (BAC) is a Bayesian model averaging method to select and adjust for confounding factors when evaluating the average causal effect of an exposure on a certain outcome. We extend the BAC method to time-to-event outcomes. Specifically, the posterior distribution of the exposure effect on a time-to-event outcome is calculated as a weighted average of posterior distributions from a number of candidate proportional hazards models, weighing each model by its ability to adjust for confounding factors. The Bayesian Information Criterion based on the partial likelihood is used to compare different models and approximate the Bayes factor. The posterior sample of the exposure effect is obtained using STAN. Performance of our method is assessed using simulation studies.

email: li.xu@uky.edu

BayesESS: AN R PACKAGE AND A WEB-BASED APPLICATION FOR QUANTIFYING THE IMPACT OF PARAMETRIC PRIORS IN BAYESIAN ANALYSIS

Jaejoon Song,* U.S. Food and Drug Administration
Jiun-Kae Jack Lee, University of Texas MD Anderson Cancer Center
Satoshi Morita, Kyoto University

Fundamental to Bayesian inference involves the specification of prior distribution and computation techniques when going from prior to posterior knowledge. While

much advances on Bayesian computations have been made recently, there is a void in software to evaluate the impact of prior distribution on posterior inference. The impact of prior on posterior is often measured by the effective samples size. This conceptualization quantifies the magnitude of influence posed by the choice of prior distribution on the posterior distribution in the unit of sample size. We introduce BayesESS, a free, open-source, comprehensive R package for quantifying the impact of parametric priors in Bayesian analysis We also introduce an accompanying web-based application for estimating and visualizing Bayesian effective sample sizes for the purpose of planning or conducting Bayesian analyses.

email: jaejoon.song@fda.hhs.gov

47. FUNCTIONAL DATA APPLICATIONS AND METHODS

ACTIVITY CLASSIFICATION USING THE SMARTPHONE GYROSCOPE AND ACCELEROMETER

Emily Huang*, Harvard T.H. Chan School of Public Health
Jukka-Pekka Onnela, Harvard T.H. Chan School of Public Health

Many medical applications use different types of activities, such as walking, sitting, or climbing stairs, as an outcome or covariate of interest. Reserachers have traditionally relied on surveys to quantify subjects' activity levels, but surveys are subjective in nature and have known limitations, such as recall bias. Smartphones provide an opportunity for objective measurement during subjects' daily lives in naturalistic settings. We explore the potential of using subjects' smartphones to estimate their levels of walking, sitting, standing, and using stairs. Focusing on the gyroscope and accelerometer sensors, we conducted a study in which participants performed various activities with one phone in their front pocket and another phone in their back pocket. We apply a modified version of the so-called movelet method to accurately detect activities and to differentiate between standing, walking, going up stairs, and going down stairs. Our results demonstrate the promise of smartphones for activity classification in naturalistic settings.

email: ehuang@hsph.harvard.edu

QUANTIFYING PHYSICAL ACTIVITY WITH SMARTPHONE ACCELEROMETRY

Josh Barback*, Harvard T.H. Chan School of Public Health

Human physical activity is temporally inhomogeneous, exhibiting both bursty periods and correlated events. Specifically, durations of active and sedentary behavior follow heavy-tailed distributions. Past research has leveraged Actigraphy-based methods to examine these distributions over periods of several weeks, with the goal of identifying features that characterize psychiatric disorders. Recent innovations in smartphone data collection permit observation of physical activity via accelerometry for much longer follow-up periods. These innovations can yield novel strategies for monitoring patient behavior and for delivering personalized interventions. However, these sensor data are subject to interruptions arising from diverse smartphone use habits. We present methods for the analysis of physical activity distributions in the presence of missingness typical of smartphone sensor observations, with applications to data from ongoing studies at Harvard Medical School teaching hospitals.

email: barback@fas.harvard.edu

ADDRESSING MISSING ACCELEROMETER DATA WITH FUNCTIONAL DATA ANALYSIS (FDA)

Patrick Hilden*, Columbia University
Jeff Goldsmith, Columbia University
Joseph Schwartz, Columbia University
Kieth Diaz, Columbia University
Ipek Ensari, Columbia University

The use of accelerometers to assess physical activity in research studies has increased in recent years. Commercial devices represent a low cost, durable, and unobtrusive option for recording physical activity over extended time periods. Recent advances have allowed a shift in observation from the laboratory setting to the day to day lives of participants. Increases in observation time has led to subsequent increases in missing data due to non-wear, and a need for ways to address this. FDA, which regards the unit of observation to be a function over time or space, has clear applications in accelerometer research. Functional principal components analysis (FPCA), which aims to approximate functional observations by a combination of primary modes of variability, provides an effective strategy for estimating functional profiles in addition to data reduction. We propose a simulation study evaluating FPCA and a recent nonnegative decomposition approach (akin to FPCA) as methods for estimating the underlying functional profiles, and subsequently imputing plausible values for missing data. Evaluation of these methods on a real-world data set will also be discussed.

email: philden@gmail.com

A FUNCTIONAL MIXED MODEL FOR SCALAR ON FUNCTION REGRESSION WITH APPLICATION TO A FUNCTIONAL MRI STUDY

Wanying Ma*, North Carolina State University
Bowen Liu, North Carolina State University
Luo Xiao, North Carolina State University
Martin A. Lindquist, Johns Hopkins Bloomberg School of Public Health

Motivated by a functional MRI study, we propose a novel functional mixed model for scalar on function regression. The model extends the standard scalar on function regression for repeated outcomes by incorporating random subject-specific functional effects. Using functional principal component analysis, the new model can be reformulated as a mixed effects model and hence can be easily fitted. A test is also proposed to assess the existence of the random subject-specific functional effects. We evaluate the performance of the model and the test via a simulation study as well as the motivating fMRI study.

email: wma9@ncsu.edu

COVARIATE-ADJUSTED REGION-REFERENCED GENERALIZED FUNCTIONAL LINEAR MODEL FOR EEG DATA

Aaron W. Scheffler*, University of California, Los Angeles
Donatello Telesca, University of California, Los Angeles
Catherine A. Sugar, University of California, Los Angeles
Shafali Jeste, University of California, Los Angeles
Abigail Dickinson, University of California, Los Angeles
Charlotte DiStefano, University of California, Los Angeles
Damla Senturk, University of California, Los Angeles

Electroencephalography (EEG) studies produce region-referenced functional data in the form of EEG signals recorded across electrodes on the scalp. In our motivating study, resting state EEG is collected on both typically developing (TD) children and children with Autism Spectrum Disorder (ASD) aged two to twelve years old. The peak alpha frequency (PAF), defined as the location of a prominent peak in the alpha frequency band of the spectral density, is an important biomarker linked to neurodevelopment and is known to shift from lower to higher frequencies as children age. To retain the most amount of information from the data, we consider the oscillations in the spectral density within the alpha band, rather than just the peak location, as a functional predictor of diagnostic status (TD vs. ASD), adjusted for chronological age. A covariate-adjusted region-referenced generalized functional linear model (CARR-GFLM) is proposed for modeling scalar outcomes from region-referenced functional predictors, which utilizes a tensor basis to estimate functional effects across a discrete regional domain while simultaneously adjusting for additional non-functional covariates, such as age.

email: ascheffler@ucla.edu

FUNCTION-ON-SCALAR QUANTILE REGRESSION WITH APPLICATION TO MASS SPECTROMETRY PROTEOMICS DATA

Yusha Liu*, Rice University
Meng Li, Rice University
Jeffrey S. Morris, University of Texas MD Anderson Cancer Center

Mass spectrometry proteomics can be used to identify potential cancer biomarkers. Existing mass spectrometry analyses utilize mean regression to detect differentially expressed proteins across groups. However, given the inter-patient heterogeneity that is a key hallmark of cancer, many biomarkers are only present at aberrant levels for a subset of cancer samples. Differences in these biomarkers might be missed by mean regression, but more easily detected by quantile-based approaches. Thus, we propose a Bayesian framework to perform quantile regression on functional responses. Our approach utilizes an asymmetric Laplace likelihood, represents the functional coefficients with basis functions, and places a global-local shrinkage prior on the basis coefficients to achieve adaptive regularization. An efficient Gibbs sampler is developed to draw posterior samples that can be used to perform Bayesian estimation and inference while accounting for multiple testing. Our framework achieves greatly improved performance over competing methods, as demonstrated by simulations. We apply this model to identify proteomic biomarkers of pancreatic cancer missed by functional mean regression.

email: yl95@rice.edu

IDENTIFICATION OF PROBLEMATIC CELL LINES FROM IN VITRO DRUG RESPONSE DATA

Farnoosh Abbas Aghababazadeh*, H. Lee Moffitt Cancer Center
Brooke L. Fridley, H. Lee Moffitt Cancer Center

Cancer cell lines (CCLs) have made a substantial contribution to cancer research by enabling the testing of chemotherapies. The cell line resources often used are the CCLE and GDSC which contain drug response data on 504 and 990 CCLs, respectively. Questions have been raised recently about the reliability of studies with CCLs. Thus, we have developed a nonlinear mixed-effects (NLME) model to determine any CCLs that tend to be overly sensitive or resistant to the majority of drugs, as these samples might represent outlier CCLs. The optimal functional form for each cell line is determined by fitting 3-parameter or 4-parameter logistic models. Then, for each cancer type and drug combination, a NLME model is fitted to the drug response data collected on the CCLs that had the same functional form. The resulting estimates for the random effects for each cell line were then standardized to allow comparison across drugs, as each drug's random effects are allowed to have a different variance. Then, we determined CCLs sensitive or resistant to the drug and summarized these findings across all cell lines and drugs.

email: Farnoosh.AbbasAghababazadeh@moffitt.org

48. MACHINE LEARNING AND STATISTICAL RELATIONAL LEARNING

USING DEEP LEARNING FOR AUTOMATED SCORING OF ANIMAL SLEEP AND WAKE STATES IN NEUROSCIENCE RESEARCH AND DEVELOPMENT

Vladimir Svetnik*, Merck & Co.
Ting-Chuan Wang, Merck & Co.
Yuting Xu, Merck & Co.

Automated scoring of animal sleep-wake states is important in neuroscience fundamental research and pharmaceutical development Traditionally, is was done using predefined features extracted from continuously measured EEG, EMG, EOG, and activity signals acquired from animals. Existing scoring systems have several drawbacks necessitating development of new systems including ones based on Deep Learning. Using these methods and large database of signals collected from the animals over many years, we have developed algorithms and software for sleep-wake scoring rodents, canines, and monkeys. Three approaches were considered and compared including deep multichannel CNN with each channel corresponding to one of the measured signals; Bagging of the pre-defined Fourier spectral features derived from the signals; and, for canines and monkeys only, Transfer Learning where the features first are learned by the multichannel CNN using human sleep-wake training data and then transferred to the sleep sleep-wake scoring in animals. We report results of the extensive analysis of the algorithm performance and recommendations for their implementation.

email: vladimir_svetnik@merck.com

MACHINE LEARNING FOR PROTEIN DESIGN

Yuting Xu*, Merck & Co.
Andy Liaw, Merck & Co.

Novel protein design is an important task in pharmaceutical research and development. Recent advances in technology enabled efficient protein design by mimicking natural evolutionary mutation, selection and amplification steps in a laboratory environment. However, due to the astronomically large number of possible polypeptide and amino acid sequences, it is difficult to explore the functionally interesting variants and it remains impossible to search for all combinations of sequences. In this work, we developed Machine Learning and Deep Learning methods for predicting properties based on their sequences. Our prediction models successfully identify the promising mutations of protein sequences in prospective time-split training and test data sets. The result indicates that the approach can potentially speed up the protein design process significantly.

email: yuting.xu@merck.com

ESTIMATING OPTIMAL TREATMENT REGIMES USING MULTIVARIATE RANDOM FORESTS

Boyi Guo*, University of Alabama at Birmingham
Ruoqing Zhu, University of Illinois at Urbana-Champaign
Hannah D. Holscher, University of Illinois at Urbana-Champaign
Loretta S. Auvil, University of Illinois at Urbana-Champaign
Michael E. Welge, University of Illinois at Urbana-Champaign
Colleen B. Bushell, University of Illinois at Urbana-Champaign
David J. Baer, Beltsville Human Nutrition Research Center
Janet A. Novotny, Beltsville Human Nutrition Research Center

With the rapid development of precision medicine in recent years, optimal treatment regime estimation has become one of the most popular topics in statistics and machine learning communities. Existing methods focus mostly on one-dimensional response. However, in complicated clinical studies, multiple endpoints are recorded and used for the evaluation of treatments. Furthermore, existing methods cannot fully utilize all information, ignoring the change of covariates induced by a treatment. To tackle these problems, we propose a tree-based model and its ensemble version for estimating optimal treatment decisions when multi-dimensional response is presenting. These models recursively partition covariate space by scanning for the best splitting rule that isolates populations whose treatment effect differs the most. Instead of relying on a pre-specified summary, our methods summarize the multi-dimensional response dynamically in each internal node of a tree or a forest, referring to the covariance structures of covariates and multi-dimensional response. Extensive simulation studies suggest that the method outperforms existing methods.

email: boyiguo1@uab.edu

MACHINE LEARNING ALGORITHMS FOR PARTIALLY SUPERVISED DATA WITH APPLICATIONS IN GROUP TESTING

Michael R. Stutz*, University of South Carolina
Zichen Ma, University of South Carolina
Joshua M. Tebbs, University of South Carolina

Group testing, or pooled testing, is a method aimed at efficient detection and prevalence estimation of a characteristic of interest, often disease, whereby individual biospecimens (e.g., blood, urine, etc.) are pooled together and tested as a whole for the presence of the characteristic. Modeling group testing data falls under the umbrella of partially supervised learning as the true characteristic statuses are latent. This makes applying readily available regression techniques particularly difficult. In this paper, we propose a framework for modeling group testing data which consists of combining machine learning techniques with the expectation-maximization algorithm. The proposed algorithms offer a novel approach to modeling group testing data, especially in light of ever more prevalent big data. Our approach can simultaneously estimate the prevalence, individual characteristic probabilities, and assay accuracy probabilities from data which arises from any group testing methodology. We compare our methods to existing methods via simulation as well as using chlamydia data obtained through group testing.

email: stutzm@email.sc.edu

ROBUST NONPARAMETRIC METHODS FOR DIFFERENCE-IN-DIFFERENCES DESIGNS

Toyya A. Pujol*, Georgia Institute of Technology
Sherri Rose, Harvard Medical School

Typical methods for difference-in-differences analyses rely on parametric statistical models that make strong assumptions about the unknown underlying functional form of the data. These assumptions are often violated in practice. Our project will extend existing statistical machine learning methods to target a difference-in-differences parameter, defined nonparametrically, while considering a larger, less restrictive nonparametric model space that makes fewer assumptions. Thus, we will develop a general statistical framework for difference-in-differences designs that allow researchers to estimate causal or statistical effect quantities using double robust machine learning while providing statistical inference. We show that the estimator remains grounded in asymptotic theory and has strong practical empirical performance. The project will apply the developed method to estimate the effects of episode-based bundle payment (EBP) on perinatal spending. We assess the commercial claims of Arkansas, a state that implemented EBP for perinatal care, and several neighboring states to determine the impact of EBP on spending.

email: pujol@gatech.edu

DETERMINING THE NUMBER OF LATENT FACTORS IN STATISTICAL MULTI-RELATIONAL MODEL

Chengchun Shi*, North Carolina State University
Wenbin Lu, North Carolina State University
Rui Song, North Carolina State University

Statistical relational learning is primarily concerned with learning and inferring relationships between entities in large-scale knowledge graphs. Nickel et al. (2011) proposed a RESCAL tensor factorization model for statistical relational learning, which achieves better or at least comparable results on common benchmark datasets when compared to other state-of-the-art methods. Given a positive integer s, RESCAL computes an s-dimensional latent vector for each entity. The latent factors can be

further used for solving relational learning tasks. The focus of this talk is to determine the number of latent factors in RESCAL. Due to the structure of RESCAL, its log-likelihood function is not concave. As a result, the corresponding maximum likelihood estimators (MLEs) may not be consistent. Nonetheless, we design a specific pseudometric, prove the consistency of the MLEs under this pseudometric and establish its rate of convergence. Based on these results, we propose a general class of information criteria and prove their model selection consistencies. Numerical examples show that our proposed information criteria have good finite sample properties.

email: cshi4@ncsu.edu

49. INTERVAL-CENSORED AND MULTIVARIATE SURVIVAL DATA

MULTIVARIATE PROPORTIONAL INTENSITY MODEL WITH RANDOM COEFFICIENTS FOR EVENT TIME DATA WITH APPLICATION TO PROCESS DATA FROM EDUCATIONAL ASSESSMENT

Hok Kan Ling*, Columbia University
Jingchen Liu, Columbia University
Zhiliang Ying, Columbia University

Motivated by the need for statistical modeling and analysis of process data, we propose a multivariate proportional intensity model with random coefficients, as such data typically consist of multi-type event time data. In an exploratory analysis on process data, a large number of possibly time-varying covariates maybe included. These covariates along with the high-dimensional counting processes often exhibit a low-dimensional structure that has meaningful interpretation. We explore such structure through specifying random coefficients in a low-dimensional space. Furthermore, to obtain a parsimonious model and to improve interpretation of parameters therein, variable selection and estimation for both fixed and random effects are developed by penalized likelihood. The computation is carried out by a stochastic EM algorithm. Simulation studies demonstrate that the proposed estimation provides an effective recovery of the true structure in both fixed and random effects. The proposed method is applied to analyzing the log-file of an item from the Programme for the International Assessment of Adult Competencies (PIAAC), where meaningful relationships are discovered.

email: hl2902@columbia.edu

METHOD FOR EVALUATING LONGITUDINAL FOLLOW-UP FREQUENCY: APPLICATION TO DEMENTIA RESEARCH

Leah H. Suttner*, University of Pennsylvania
Sharon X. Xie, University of Pennsylvania

Current practice of longitudinal follow-up frequency in outpatient clinical research is mainly based on experience, tradition, and availability of resources. Previous methods for designing follow-up times require parametric assumptions about the hazards for an event. There is a need to develop robust, easy to implement, quantitative procedures for justifying the appropriateness of follow-up frequency. We propose a novel method to evaluate follow-up frequency by assessing the impact of right-endpoint imputation of interval-censored data in longitudinal studies. Specifically, we evaluate the bias in estimating hazard ratios using Cox models under various follow-up schedules. Our simulation-based procedure applies the schedules to generated data resembling the survival curve of historical data. Using this method, we evaluate the current follow-up of Parkinson's disease (PD) patients at the University of Pennsylvania Parkinson's disease Research Center. However, the method can be applied to any research area with sufficient historical data for appropriate data generation. To allow clinical investigators to implement this method, we provide a Shiny web application.

email: lsutt@pennmedicine.upenn.edu

COPULA-BASED SIEVE SEMIPARAMETRIC TRANSFORMATION MODEL FOR BIVARIATE INTERVAL-CENSORED DATA

Tao Sun*, University of Pittsburgh
Wei Chen, University of Pittsburgh
Ying Ding, University of Pittsburgh

This research is motivated by discovering genetic causes for the progression of a bilateral eye disease, Age-related Macular Degeneration (AMD), of which the primary outcomes, progression times to late-AMD, are bivariate and interval-censored. We develop a copula-based semiparametric approach for modeling and testing bivariate interval- censored data. Specifically, the joint likelihood is modeled through a two-parameter Archimedean copula, which can flexibly characterize the dependence between two margins. The marginal distributions are modeled through a semiparametric transformation model using sieves, with the proportional hazards or odds model being a special case. We propose a sieve maximum likelihood estimation procedure and develop a generalized score test for testing the regression parameter(s). For the proposed sieve estimators of finite-dimensional parameters, we establish their asymptotic normality and efficiency. Extensive simulations are conducted to evaluate the performance of the proposed method in finite samples. Finally, we apply our method to a genome-wide analysis of AMD progression, to identify susceptible risk variants for the disease progression.

email: tao.sun@pitt.edu

BAYESIAN REGRESSION ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED FAILURE TIME DATA UNDER THE NORMAL FRAILTY PROBIT MODEL

Yifan Zhang*, University of South Carolina
Lianming Wang, University of South Carolina

Interval-censored data naturally arise in many epidemiological, social- behavioral, and medical studies, in which subjects are examined multiple times and the failure times of interest are not observed exactly, but fall within some intervals. A new frailty probit model is proposed for the regression analysis of multivariate interval-censored data, and this model allows explicit form of the pairwise statistical association among the failure times. An efficient Bayesian estimation approach is proposed under this model and allow joint estimation of regression parameters and other secondary parameters. The proposed method is evaluated by extensive simulation studies and illustrated by a real-life application.

email: zhang374@email.sc.edu

A PROPORTIONAL HAZARDS MODEL FOR INTERVAL-CENSORED DATA SUBJECT TO INSTANTANEOUS FAILURES

Prabhashi W. Withana Gamage*, James Madison University
Monica Chaudari, University of North Carolina, Chapel Hill
Christopher S. McMahan, Clemson University
Edwin H. Kim, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

The proportional hazards (PH) model is arguably one of the most popular models used to analyze time to event data. In many studies, the event time is not directly observed but is known relative to examination times. Further, in some studies the observed data also consists of instantaneous failures; i.e., the event times for several study units coincide exactly with the time at which the study begins. This work focuses on developing a mixture model, under the PH assumptions, which can be used to analyze interval-censored data subject to instantaneous failures. To allow for modeling flexibility, two methods of estimating the cumulative baseline hazard function are proposed; a fully parametric and a monotone spline representation. Through a novel data augmentation procedure, an expectation-maximization algorithm is developed to complete model fitting. Through simulation studies the proposed approach is shown to provide reliable estimation and inference. The motivation for this work arises from a randomized clinical trial aimed at assessing the effectiveness of a new peanut allergen treatment in attaining sustained unresponsiveness in children.

email: withanpw@jmu.edu

50. UNDERSTANDING THE COMPLEXITY AND INTEGRITY OF CLINICAL TRIAL DATA

SIMULATING REALISTIC CLINICAL TRIAL DATA

Naji Younes*, The George Washington University

This talk explores the difficulties that arise when simulating realistic datasets representing the baseline characteristics and longitudinal experience of participants in a clinical trial. The variables in such datasets are related to each other in complex ways, and constrained by biological considerations and the logistics of the clinical trial. We'll discuss statistical and computational issues, and the implications these have on the analysis of data.

email: naji@bsc.gwu.edu

THE P-VALUE REQUIRES CONTEXT AND CORRECT INTERPRETATION IN CLINICAL TRIALS

Rebecca Betensky*, NYU College of Global Public Health

In this talk, I will present two in-depth considerations of the p-value and its potential pitfalls. In conjunction with the design and context of the study, such as sample size and the minimum meaningful effect size, which are inputs to the calculation of confidence limits for measures of effect, the p-value may be informative about the effect of interest and/or about the null. However, absolute thresholds for the p-value do not render it meaningful with regard to a positive or null effect; the thresholds depend on sample size and effect size. This understanding expands on the ASA statement (Wasserstein and Lazar, 2016), which enumerates truisms about the p-value, but does not provide guidance regarding best uses of the p-value, and provides nuance to the simple stringent threshold suggested by Benjamin et al. (2017). I also describe the potential problems associated with using baseline p-values from a clinical trial to assess the validity of the randomization due to correlation among those p-values.

email: rebecca.betensky@nyu.edu

DETECTING FRAUDULENT BASELINE DATA IN CLINICAL TRIALS

Michael A. Proschan*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Pamela A. Shaw, University of Pennsylvania Perelman School of Medicine

The first table in many articles reporting the results of a randomized clinical trial compares baseline factors across arms. Results that appear inconsistent with chance trigger suspicion, and in one case, accusation and confirmation of data falsification. We confirm theoretically results of simulation analyses showing that inconsistency with chance is extremely difficult to establish in the absence of any information about correlations between baseline covariates.

email: ProschaM@mail.nih.gov

51. REPLICABILITY IN BIG DATA PRECISION MEDICINE

REPRODUCIBILITY AND HETEROGENEITY IN META-ANALYSIS AND REPLICATION OF TRANSCRIPTOMIC STUDIES

George Tseng*, University of Pittsburgh

Although meta-analysis aims to combine information from multiple studies to increase statistical power, heterogeneities are prevalent in genomic meta-analysis or replication studies. Such inconsistency across cohorts brings difficulties in validating gene signatures or prediction models. In the first part of the talk, we will review several methods we have developed for quality control, biomarker detection and machine learning for omics meta-analysis. In the second part, we will present a new approach of using top-scoring-pair method for robust feature selection and machine learning in the meta-analytic framework and how it improves reproducibility in machine learning of omics data.

email: ctseng@pitt.edu

CURRENT TOPICS IN MULTI-STUDY LEARNING

Prasad Patil*, Harvard T.H. Chan School of Public Health and Dana-Farber Cancer Institute
Giovanni Parmigiani, Harvard T.H. Chan School of Public Health and Dana-Farber Cancer Institute

We examine the potential for improvement of replicability when a predictor is

trained using multiple studies' worth of data. It has been shown that a predictor trained on any one study may exhibit variable performance when applied in others. The inter-study heterogeneity that drives this variability in performance can be attributed to differences in sampling, covariate shift, differing measurement techniques, as well as unobserved confounding. In the multi-study setting, we have the ability to both account for some of this heterogeneity and to use it to our advantage for increasing generalizability. Our preliminary findings have suggested that the choice of learner and the amount of inter-study heterogeneity have significant impacts on the success of any multi-study learning strategy. We will expand upon these findings and discuss other practical issues, including statistical approaches for deciding when to combine studies and strategies for weighted ensembling of predictors trained in different studies.

email: ppatil@jimmy.harvard.edu

A STATISTICAL FRAMEWORK FOR MEASURING REPLICABILITY AND REPRODUCIBILITY OF HIGH-THROUGHPUT DATA FROM MULTIPLE LABS

Qunhua Li*, The Pennsylvania State University
Monia Ranalli, Tor Vergata University, Rome, Italy
Yafei Lyu, University of Pennsylvania

High-throughput technologies play an important role in modern biological studies. Nowadays, there are a large amount of high-throughput sequencing data in the public domain. Many measured the same biological questions, but were generated from different labs or different studies. It is of great interest to evaluate both the reproducibility of the findings within the same source and the replicability of the findings from different sources. We propose nestedIDR, a novel nested copula mixture model to measure the reproducibility and replicability of the findings. This method takes account of the heterogeneity of the replicate samples from different sources, measuring both reproducibility and replicability simultaneously. Our applications on ChIP-seq and RNA-seq data show that it effectively depicts the source of the heterogeneity and rescues the signals that are not reproducible within a lab but replicated by different sources.

email: qunhua.li@psu.edu

MODELING BETWEEN-STUDY HETEROGENEITY FOR IMPROVED REPLICABILITY IN GENE SIGNATURE SELECTION AND CLINICAL PREDICTION

Naim Rashid*, University of North Carolina, Chapel Hill
Quefeng Li, University of North Carolina, Chapel Hill
Jen Jen Yeh, University of North Carolina, Chapel Hill
Joseph Ibrahim, University of North Carolina, Chapel Hill

In the genomic era, the identification of gene signatures associated with disease is of significant interest. Such signatures are often used to predict clinical outcomes in new patients and aid clinical decision-making. However, recent studies have shown that gene signatures are often not replicable. This occurrence has practical implications in the generalizability and clinical applicability of such signatures. To improve replicability, we introduce a novel approach to select gene signatures from multiple data sets whose effects are consistently non-zero by accounting for between-study heterogeneity. We build our model upon some platform-robust quantities, enabling integration over different platforms of genomic data. A high dimensional penalized GLMM is used to select gene signatures and address data heterogeneity. We compare our method to two commonly used strategies ignoring between-study heterogeneity, and show that these strategies have inferior performance in predicting outcome in new studies. Lastly, we motivate our method through a case study subtyping pancreatic cancer patients from four studies using different gene expression platforms.

email: naim@unc.edu

52. COMPUTATIONALLY-INTENSIVE BAYESIAN TECHNIQUES FOR BIOMEDICAL DATA: RECENT ADVANCES

FINDING AND LEVERAGING STRUCTURE WITH BAYESIAN DECISION TREE ENSEMBLES

Antonio R. Linero*, Florida State University
Junliang Du, Florida State University

Analysis of complex or high-dimensional datasets is often aided when the data is structured; such structures include sparsity and low-order interaction structures, as well as a-priori known graphical structures obtained from external data sources. In this talk, we present strategies for finding and leveraging structural information in order to boost the performance of Bayesian decision tree ensembles. We focus on three problems: (i) detection of low-order interactions in datasets; (ii) use of grouping information, similar to the group and overlapping-group lasso; and (iii) graphical structures which encode a-priori known relationships between predictors. The methods we develop provide powerful, nonparametric, alternatives to existing frequentist and Bayesian approaches which have focused almost exclusively on linear models, with implementation of our approaches requiring minor modifications to existing algorithms. We provide simulation evidence for the benefits of our proposed approaches, and apply the methodology to several datasets.

email: arlinero@stat.fsu.edu

BAYESIAN ESTIMATION OF INDIVIDUALIZED TREATMENT-RESPONSE CURVES IN POPULATIONS WITH HETEROGENEOUS TREATMENT EFFECTS

Yanxun Xu*, Johns Hopkins University
Yanbo Xu, Georgia Tech University
Suchi Saria, Johns Hopkins University

Estimating individual treatment effects is crucial for individualized or precision medicine. We use non-experimental data; we model heterogenous treatment effects in the studied population and provide a Bayesian estimator of the individual treatment response. More specifically, we develop a novel Bayesian nonparametric (BNP) method that leverages the G-computation formula to adjust for time-varying confounding in observational data, and it flexibly models sequential data to provide posterior inference over the treatment response at both group level and individual level. On a challenging dataset containing time series from patients admitted to intensive care unit (ICU), our approach reveals that these patients have heterogenous

responses to the treatments used in managing kidney function. We also show that on held out data the resulting predicted outcome in response to treatment (or no treatment) is more accurate than alternative approaches.

email: yanxunxu.stat@gmail.com

MONOTONE SINGLE-INDEX MODELS FOR HIGHLY SKEWED RESPONSE

Debajyoti Sinha*, Florida State University
Kumaresh Dhara, University of Florida
Bradley Hupf, Florida State University
Greg Hajcak, Florida State University

Single-index models are practical, useful tools for modeling and analyzing many clinical and mental-health studies with complex non-linear covariate effects on a highly skewed response variable. We propose Bayesian methods for estimating monotone single-index models of quantiles using the suitable basis representation of the monotone link function. The monotonicity of the unknown link function offers a clinical interpretation of the index, along with the relative importance of the components of the index. To ease the computational complexity of the Bayesian analysis, we develop a novel and efficient Metropolis-Hastings step to sample from the conditional posterior distribution of the index parameters. These methodologies and their advantages over existing methods are illustrated via simulation studies and analysis of a study of the risks of depression among adolescent girls.

email: sinhad@stat.fsu.edu

A GRAPHICAL MODEL FOR SKEWED MATRIX-VARIATE NON-RANDOMLY MISSING DATA

Dipankar Bandyopadhyay*, Virginia Commonwealth University
Lin Zhang, University of Minnesota

Periodontal disease (PD) studies collect relevant bio-markers, such as clinical attachment level (CAL) and the probed pocket depth (PPD), at tooth-sites, along with various other demographic and biological risk factors. Although routine cross-sectional evaluation under a linear mixed model (LMM) framework with underlying normality assumptions are popular, a careful investigation reveals considerable non-normality manifested in the random terms in the form of skewness and tail behavior. In addition, PD progression is hypothesized to be spatially-referenced. To mitigate these complexities, we consider a matrix-variate skew-t formulation of the LMM with a Markov graphical embedding for modeling bivariate (PPD & CAL) responses, with the non-randomly missing responses imputed via a latent probit regression. Our hierarchical Bayesian framework addresses the aforementioned complexities within an unified paradigm and provides seamless stimation of model parameters Using both synthetic and a clinical data assessing PD status, we demonstrate the advantages of our proposal over known alternatives.

email: dbandyop@vcu.edu

53. MULTIVARIATE FUNCTIONAL DATA ANALYSIS WITH MEDICAL APPLICATIONS

DIMENSION REDUCTION FOR FUNCTIONAL DATA BASED ON WEAK CONDITIONAL MOMENTS

Bing Li*, The Pennsylvania State University
Jun Song, University of North Carolina, Charlotte

We develop a general theory and estimation methods for functional linear sufficient dimension reduction, where both the predictor and the response can be vectors of functions. Unlike the existing dimension reduction methods, our approach does not rely on the estimation of conditional mean and conditional variance. Instead, it is based on a new statistical construction --- the weak conditional expectation. Weak conditional expectation is a generalization of conditional expectation. Its key advantage is to replace the projection on to an L2-space --- which defines conditional expectation --- by projection on to an arbitrary Hilbert space, while still maintaining the unbiasedness of the related dimension reduction methods. This flexibility is particularly important for functional data, because attempting to estimate a full-fledged conditional mean or conditional variance by slicing or smoothing over the space of vector-valued functions may be inefficient due to the curse of dimensionality. We evaluated the performances of the our new methods by simulation and in several applied settings.

email: bxl9@psu.edu

ALIGNMENT OF fMRI TIME-SERIES AND FUNCTIONAL CONNECTIVITY

Jane-Ling Wang*, University of California, Davis
Chun-Jui Chen, University of California, Davis

Due to technology advance, spatially indexed objects are commonly observed across different scientific disciplines. Such object data are typically high-dimensional and pose great challenges to scientists due to the curse of high-dimensionality. While sparsity is commonly adopted as an assumption in high-dimensional settings, its validity is difficulty to verify. We propose a new approach for spatially indexed object data by mapping their spatial locations to a targeted one-dimensional interval so objects that are similar are placed near each other on the new target space. The proposed alignment provides a visualization tool to view these complex object data. Moreover, the aligned data often exhibit certain level of smoothness and can be handled by approaches designed for functional data. We demonstrate how to implement such an alignment for fMRI time series and propose a new concept of path length to study functional connectivity, in addition to a new community detection method. The proposed methods are illustrated by simulations and on a study of Alzheimer's disease.

email: janelwang@ucdavis.edu

FUNCTIONAL MARGINAL STRUCTURAL MODELS FOR TIME-VARYING CONFOUNDING OF MOOD ASSESSMENTS

Haochang Shou*, University of Pennsylvania

The increasing availability of the dense measure of various biosignals ihas provided opportunities for us to learn the integrative relationship of the multi-domain biological systems, yet also poses challenges for statistical analysis. For example, in psychiatric studies, wearable sensors such as accelerometers or smart phones are often built in to record the participants' physical activity pobjectively and continuously over days. Meanwhile, participants are also expected to answer electronic surveys about their mood, sleep and behaviors several times a day. While the time-varying effects of mood conditions (e.g. sadness, anxiety) on endpoints like episode onset are of interest, the activity intensities might potentially affect mood at the following point, and hence confound the association between the main exposure variable and endpoint. Here we propose a marginal structural model type framework for functional data such as continuous daily physical activity profiles and use inverse probability weighting to correct for potential biasness induced by the time-dependent confounding effects of physical activities on mood.

email: hshou@pennmedicine.upenn.edu

FINDING BIOMARKERS FOR CHILDHOOD OBESITY USING FUNCTIONAL DATA ANALYSIS

Matthew Reimherr*, The Pennsylvania State University
Sara Craig, The Pennsylvania State University
Kateryna Makova, The Pennsylvania State University
Francesca Chiaromonte, The Pennsylvania State University
Alice Parodi, Milano di Politecnico
Junli Lin, The Pennsylvania State University
Ana Kenney, The Pennsylvania State University

In this talk I will present recent work concerning the analysis of longitudinal childhood growth trajectories using functional data analysis. We explore both the microbiome and the genome for biomarkers that put children at greater risk for obesity. We discuss tools for both variable selection and parameter estimation when the outcome is functional and one has a larger number of scalar predictors.

email: mreimherr@psu.edu

54. METHODS FOR EXAMINING HEALTH EFFECTS OF EXPOSURE TO THE WORLD TRADE CENTER ATTACK AND BUILDING COLLAPSE

MODELING COMORBID MENTAL AND MEDICAL OUTCOMES VIA LATENT CLASS REGRESSION

Yongzhao Shao*, New York University School of Medicine

There is a lack of literature on modeling comorbidity in mental and physical health conditions from human civilian populations after environmental exposures. The comorbidity in mental and physical health disorders is prevalent among the residents and responders exposed to the environmental catastrophe after the World Trade Center (WTC) towers collapse. We present latent class regression models to identify the joint comorbid disease patterns and clusters as well as to characterize the exposure and other factors underlying the mental and physical health clusters. Variable selection methods and computing algorithms for implementing the proposed latent class regression models are evaluated using simulation studies and illustrated using the WTC data sets.

email: yongzhao.shao@nyumc.org

HANDGRIP STRENGTH OF WORLD TRADE CENTER RESPONDERS: THE LONG-TERM ROLE OF RE-EXPERIENCING TRAUMATIC EVENTS

Sean A. Clouston*, Stony Brook University
Peifen Kuan, Stony Brook University
Soumyadeep Mukherjee, Rhode Island College
Roman Kotov, Stony Brook University
Evelyn J. Bromet, Stony Brook University
Benjamin J. Luft, Stony Brook University

Handgrip strength (HGS), a measure of muscle strength, is a well known biomarker of aging. This study examined associations between posttraumatic stress disorder (PTSD) and HGS in World Trade Center responders. HGS was assessed using a computer-assisted hand dynamometer to a consecutive sample of men and women (N=2,023) who participated in rescue and recovery efforts following the 9/11/2001 attacks and attended monitoring efforts in Long Island, NY. PTSD symptom severity was assessed using the PTSD specific-trauma checklist. General linear models were used to derive measures of HGS and to examine associations between PTSD and HGS. The assessed sample was at midlife, and 91.3% were men. HGS was lower in older responders. HGS of those with probable PTSD was lower than among responders without PTSD. Subdomain analyses of PTSD symptoms revealed that re-experiencing symptoms at enrollment ($p<0.001$) and contemporary avoidance symptoms ($p=0.005$) were associated with lower HGS. The current study therefore suggests that PTSD may be associated with weaker muscle strength and more rapid aging. Future clinical efforts may be needed to improve physical capability following trauma.

email: sean.clouston@stonybrookmedicine.edu

CANCER LATENCY AFTER ENVIRONMENTAL EXPOSURE: A CHANGE POINT APPROACH

Charles B. Hall*, Albert Einstein College of Medicine

There is surprisingly little data from human populations on the latency of cancer incidence after exposure to environmental carcinogens. The relatively short (<1 year) exposure to the environmental catastrophe after the World Trade Center building collapse followed by long term follow-up of exposed workers and other survivors offers a unique opportunity to study this important scientific question. Change point models have long been used in cancer epidemiology research, but semiparametric survival models are typically nonlinear in the time and risk factor parameters, making partial likelihood methods impossible to use in this context. We present a fully parametric piecewise exponential model that overcomes these issues. The model is easilly fit using standard software using Bayesian or frequentist approaches and we

present results from preliminary analyses on cancer incidence based on linkages to multiple state cancer registries.

email: charles.hall@einstein.yu.edu

55. REGRESSION, MEDIATION, AND GRAPHICAL MODELING TECHNIQUES FOR MICROBIOME DATA

A FRAMEWORK FOR ANALYSIS OF MICROBIOME DATA WITH RESPECT TO MULTIVARIATE CLINICAL INSTRUMENTS

Alexander Alekseyenko*, Medical University of South Carolina

To provide convenient actionable clinical guidelines multi-dimensional clinical instruments are often converted aggregated summary scores, such as SLEDAI (Systemic Lupus Erythematosus Disease Activity Index). Following the clinical use case, these summary scores are also used in research to draw associations with study variables, such as the composition of the microbiome. The downside of this approach is that the fine-grained details embedded in these instruments are lost in favor of having a univariate score, which in itself may have little relevance to biomedical processes involved. Using distance-based energy statistics, I demonstrate the ability to model relationships between multivariate microbiome data and multivariate clinical responses. This allows to step beyond analyses of arbitrary aggregate summary scores with respect to clinical survey instruments. I demonstrate the application of this approach in several datasets.

email: alekseye@musc.edu

DISENTANGLING MICROBIAL ASSOCIATIONS VIA LATENT VARIABLE GRAPHICAL MODELS

Zachary D. Kurtz*, Lodo Therapeutics
Christian L. Mueller, Flat Iron Institute, Simons Foundation

Inferring associations from counts of microbial marker genes is a key task in microbiome science for the prediction of ecological and functional interactions. Normalizing or rarefying count data is a routine processing step to adjust for sampling difference, yet doing so introduces well-known compositional biases to statistical association measures. Additionally, associations based on pairwise correlation or conditional independence do not account for unobserved covariates, such as technical artifacts or environment factors. We address these concerns with the addition of a latent variable graphical model selection routine to the Sparse InversE Covariance estimation for Ecological Association Inference (SPIEC-EASI) pipeline. Under mild assumptions and in the noiseless setting, we can achieve - in theory and empirically - near exact recovery of the model support regardless of the number of compositional features. Further, we show that real data networks inferred via latent variable graphical models are more consistent and biologically interpretable than other graphical model frameworks, even when adjusted for sampling biases.

email: zdkurtz@gmail.com

ROBUST REGRESSION WITH COMPOSITIONAL COVARIATES

Aditya Kumar Mishra*, Flatiron Institute, Simons Foundation
Christian L. Mueller, Flatiron Institute, Simons Foundation

Microbiome regulates various metabolic function in human, and biogeochemical function in the marine ecosystem. With the large-scale efforts in 16S ribosomal RNA sequencing in the microbial study, we have relative abundance/compositional data of the group of microbial taxa at different taxonomic levels. We investigate the dependency of a phenotype/response like metabolite on these large dimensional compositional covariates. The problem is challenging especially in presence of either outlier or leveraged observations. In order to be robust, we propose an additive log-contrast model with the mean shift. The overparameterized model is estimated via penalized regression approach with regularization enforcing sparsity in mean shift and covariates parameters. We have demonstrated the efficacy of the approach using various simulation studies and an application relating body mass index to human gut microbiome data.

email: amishra@flatironinstitute.org

MEDIATION ANALYSIS IN INVESTIGATING THE ROLE OF MICROBIOME IN HUMAN HEALTH

Lingling An*, University of Arizona
Kyle Carter, University of Arizona
Meng Lu, University of Arizona

Host gene expression in cooperation with the microbiome has been discovered to play a large role in disease progression and response. In particular, changes in host gene expression may have a marked impact on bacterial species diversity and abundance. A popular method for capturing these interactions in disease classification models is to use regression mediation modeling, which maintains strong assumptions about the distribution and association of parameters. We propose a nonparametric approach for selecting significant mediating variables for models with high dimensional exposures and mediators. A simulation study shows improved performance compared to traditional nonparametric mediation methods.

email: anling@email.arizona.edu

56. SPEED POSTERS: EHR DATA, EPIDEMIOLOGY, PERSONALIZED MEDICINE, CLINICAL TRIALS

56a. INVITED SPEED POSTER: COMBINING INVERSE-PROBABILITY WEIGHTING AND MULTIPLE IMPUTATION TO ADJUST FOR SELECTION BIAS DUE TO MISSING DATA IN EHR-BASED RESEARCH

Sebastien Haneuse*, Harvard T.H. Chan School of Public Health
Tanayott Thaweethai, Harvard T.H. Chan School of Public Health

Among the many potential threats to validity in EHR-based studies, selection bias due to missing data is prominent. Existing missing data methods, however, such as inverse-probability weighting (IPW) and multiple imputation (MI), typically fail to acknowledge the complexity of EHR data. To resolve this, Haneuse et al (2016)

proposed to modularize the data provenance into a series of sub-mechanisms, each representing a clinical "decision". Based on this we develop a general and scalable framework for estimation and inference for regression models that permits the use of IPW and/or MI to tackle each of the sub-mechanisms in a single analysis. We refer to this as a "blended analysis strategy". Simulations show that naive use of standard methods may result in bias; that the proposed estimators have good small-sample properties; and, that a bias-variance trade-off may manifest as researchers consider how to handle missing data. The proposed methods are illustrated with data from a multi-site EHR-based study of the effect of bariatric surgery on BMI.

email: shaneuse@hsph.harvard.edu

56b. INVITED SPEED POSTER: METHODS TO UTILIZE LONGITUDINAL EHR DATA TO INVESTIGATE WHETHER MOVING TO A DIFFERENT BUILT ENVIRONMENT AFFECTS HEALTH

Jennifer F. Bobb*, Kaiser Permanente
Andrea J. Cook, Kaiser Permanente

Large multi-decade healthcare databases of longitudinal patient electronic health records (EHRs) have been used to conduct public health research. EHR data can be enhanced by being linked to other databases that provide augmented information such as the built environment that the patient lives in (e.g. neighborhood walkability, number of parks, property value, and crime statistics). Further, patients move over time into potentially different built environments yielding time-varying predictors of interest, which allows for estimation of causal association tied to health outcomes (e.g. weight) and changes in the built environment. Analyzing this type of data is complicated due to numerous factors including multi-level correlated data (spatial and longitudinal), missing data, and estimating a time-varying predictor with appropriate lag time until expected outcome response. In this talk we will provide an example of a study that links EHR data to the built environment in Washington State. We will detail different statistical approaches developed to handle the challenges presented when analyzing this type of data.

email: jennifer.f.bobb@kp.org

56c. PRAGMATIC EVALUATION OF RELATIVE RISK MODELS IN PheWAS ANALYSIS

Ya-Chen Lin*, Vanderbilt University
Siwei Zhang, Vanderbilt University Medical Center
Lisa Bastarache, Vanderbilt University Medical Center
Todd Edwards, Vanderbilt University Medical Center
Jill M. Pulley, Vanderbilt University
Joshua C. Denny, Vanderbilt University Medical Center
Hakmook Kang, Vanderbilt University
Yaomin Xu, Vanderbilt University

The phenome-wide association study (PheWAS) has become widely used for efficient, high-throughput evaluation of the relationship between a genetic factor and a collection of clinical diagnoses, typically extracted from a DNA biobank linked with electronic health records (EHR). The case-control analysis has been the standard choice for its efficiency in rare disease analysis and power to identify risk factors. However, the clinical diagnoses in EHR are often inaccurate and may cause bias. Here, we hypothesize that, relative risk (RR) is an alternative strategy that may overcome some of the limitations by case-control PheWAS analysis. Simulation with varying modeling parameters will be provided to evaluate the current RR-based methods including Log-binomial, Poisson, Cox model by its mean square error and validity of the estimates. In particular, we are interested in comparing the robustness of the method when phenotypical data are noisy and when ranking consistency in prioritizing both genetic variants and disease phenotypes. The best performance model will be applied to the selected SNPs whose disease associations are well-known and compared with the case-control analysis.

email: ya-chen.lin.1@vanderbilt.edu

56d. OPERATING CHARACTERISTICS OF BAYESIAN JOINT BENEFIT-RISK COPULA MODELS

Nathan T. James*, Vanderbilt University
Frank E. Harrell, Jr., Vanderbilt University

To receive regulatory approval, the benefits of a medical intervention must outweigh its risks. While clinical trials are designed to evaluate both efficacy and safety, outcomes for these components are often evaluated with separate models. Bayesian copula models provide a flexible, interpretable approach to joint modeling of multivariate benefit and risk outcomes by separating the modeling of each marginal outcome from the copula model that specifies the dependency between outcomes. Copula models can also be used in studies with multiple endpoints to evaluate several efficacy outcomes simultaneously. We extend the recent work of Costa and Drury (2018) by performing simulation studies to evaluate the operating characteristics (power and probability of misleading evidence) of the joint copula modeling approach compared to a joint generalized linear mixed model (GLMM) and estimation using separate models for benefit and risk. Reproducible code is provided to enable researchers to explore copula models with alternative assumptions and design characteristics.

email: nj1154@gmail.com

56e. HETEROGENEITY ASSESSMENT OF TREATMENT EFFECT AMONG SUBPOPULATIONS IN BASKET TRIALS

Ryo Sadachi*, The University of Tokyo, Japan
Akihiro Hirakawa, The University of Tokyo, Japan

Assessing heterogeneity of treatment effect among multiple subpopulations in basket trials is challenging. A Bayesian hierarchical modeling (BHM) has an attractive feature of stabilizing the estimate of treatment effect in each subpopulation, but bears the risk of too much shrinkage of treatment effect. Exchangeability-nonexchangeability (EXNEX) approach that allows each subpopulation-specific parameter to be exchangeable with other similar subpopulation parameters or nonexchangeable with any of them is a better alternative; however, the specifications of the number of EX and NEX components are required. We develop a new method for assessing the heterogeneity of treatment effect and estimating treatment effect in each subpopulation. The proposed method quantifies the similarity of treatment effect between two subpopulations based on the Kullback-Leibler (or Jensen-Shannon) divergences and

groups the subpopulations by applying an agglomerative hierarchical clustering to their similarities. We examine the utility of the proposed method through the simulation studies comparing with the BHM and EXNEX approaches.

email: ryosadachi-stat@g.ecc.u-tokyo.ac.jp

56f. EXTENSIVE COMPARISONS OF THE INTERVAL-BASED PHASE I DESIGN WITH 3+3 DESIGN FOR THE TRIALS WITH 3 OR 4 DOSE LEVELS

Jongphil Kim*, H. Lee Moffitt Cancer Center and University of South Florida

The paper focuses on the extensive comparisons of two interval-based phase I design methods, the Bayesian Optimal Interval (BOIN) design and the modified toxicity probability interval design-2 (mTPI-2), with the standard 3+3 design in situations in which the number of doses to be tested is small. These interval-based designs were recently developed and their operating characteristics were well evaluated for the cases in which the number of doses to be tested is 6 and 8. In this manuscript, the dose-escalation and de-escalation rule of the interval-based designs is slightly modified in order to be more suitable in practice and their operating characteristics will be compared with the 3+3 design for the trials with 3 or 4 dose levels. In addition, the impact of the sample size for the interval-based designs is investigated.

email: jongphil.kim@moffitt.org

56g. COMBINING EVIDENCE FROM RANDOMIZED CLINICAL TRIALS ACROSS OUTBREAKS

Natalie E. Dean*, University of Florida
Victor De Gruttola, Harvard University

For emerging infectious diseases that cause outbreaks of unpredictable size and duration, it may not be possible to accrue sufficient evidence to reliably evaluate the efficacy or effectiveness of an intervention during a single outbreak. We describe three strategies for combining randomized evidence on the efficacy or effectiveness of an intervention across outbreaks. The first makes use of a master trial that starts, stops, and restarts with each successive outbreak until sufficient information has been accumulated as specified in the trial design. The second strategy combines data across multiple underpowered trials in a blinded fashion. The third strategy is a prospective meta-analysis. We outline considerations for these approaches that permit accumulation of trial evidence across small outbreaks to guide the treatment and prevention of infectious diseases. This work is presented on behalf of the participants of the World Health Organization R&D Blueprint workplan for designing clinical trials in Public Health Emergencies.

email: nataliedean@ufl.edu

56h. CONDITIONAL QUANTILE INFERENCE WITH ZERO-INFLATED OUTCOMES

Wodan Ling*, Columbia University
Bin Cheng, Columbia University
Ying Wei, Columbia University
Ken Cheung, Columbia University

Zero-inflated outcomes are common in epidemiological studies. The phenotypes of atherosclerotic plaque burden in Northern Manhattan Study that examines the risk factors of stroke take non-negative values, with a point mass at zero. To comprehensively analyze the associations between risk factors and those phenotypes, quantile regression is a promising tool, which is an alternative of mean-based regressions, but robust to heavy-tailed distributions and more informative in providing analysis on different conditional quantiles of the outcome. However, applying quantile regression directly to zero-inflated outcomes is problematic, because the associations between covariates and quantiles of the outcome are non-linear, as long as the proportion of zero's changes according to the covariates. Therefore, we proposed a compound quantile regression framework to flexibly accommodate such heterogeneity caused by zero-inflation. In simulation and real data studies, we confirmed it provided more accurate estimation and prediction, achieved better goodness-of-fit and obtained more significant covariate effect than existing approaches.

email: wl2459@columbia.edu

56i. NONLINEAR MIXTURE MODEL FOR MODELING TRAJECTORIES OF ORDINAL MARKERS IN NEUROLOGICAL DISORDERS

Qinxia Wang*, Columbia University
Ming Sun, Columbia University
Yuanjia Wang, Columbia University

Current diagnosis of neurological disorders often relies on late-stage symptoms. Recent studies show that there is relationship between underlying disease progression and biological markers that can assist early diagnosis. We propose a nonlinear mixture model to investigate the marker trajectories, allowing for subject-specific inflection points indicating disease severity. Specifically, we focus on markers with ordinal outcomes for which higher values imply more severe impairment. The latent binary variable in the mixture model indicates disease resilience. If a subject is susceptible, we assume that he will have ordinal symptoms following an adjacent category logistic model. The odds of disease comparing adjacent categories depends on a subject's baseline measures and a subject-specific vulnerability score shared among markers. Model parameters are estimated using EM algorithm. We conduct simulation studies to demonstrate validity of the proposed model. Lastly, we apply our method to estimate the effect of personal characteristics on the marker trajectories in Parkinson's Progression Markers Initiative, and show utility to aid early personalized diagnostic decisions.

email: qw2223@cumc.columbia.edu

56j. AN AUGMENTED SURVIVAL ANALYSIS METHOD FOR INTERVAL CENSORED AND MIS-MEASURED OUTCOMES

Chongliang Luo*, University of Pennsylvania
Yong Chen, University of Pennsylvania

Survival analyses with interval censored outcomes are commonly seen in clinical and epidemiological studies, where the patients are only examined periodically and the event or failure of interest is known only to occur within a certain interval. Often the examination procedure is also subject to mis-measurement error. For example, in Electronic health records (EHR) based association study, phenotypes of patients

are derived from a high-throughput phenotyping algorithm, whereas chart reviews (deemed as a gold standard for the true phenotype) are available only for a small subset of patients. We provide a method that jointly use the error-prone outcomes of all patients and true outcomes of a validation sample to achieve bias reduction and also efficiency improvement. We presented simulation and real data examples to compare our method with other available methods.

email: luocl3009@gmail.com

56k. TRANSFORMATION OF ACTIVITY COUNTS FROM MULTIPLE ACTIVITY MONITORING DEVICES USING LATENT CORRELATION

Jordan Johns*, Johns Hopkins School of Public Health
Vadim Zipunnikov, Johns Hopkins School of Public Health
Ciprian Crainiceanu, Johns Hopkins School of Public Health

The technology surrounding activity monitoring is constantly changing with new devices and applications continually being introduced and updated. Each activity monitoring device has a unique, proprietary formula for transforming sub-second measures of acceleration to summary counts of activity over the course of a minute, hour, or day. Due to these differences and their proprietary nature, it can be difficult to determine how best to aggregate activity information from different activity monitoring devices. One example comes from the Baltimore longitudinal study of aging where scientific and technological developments caused the original objectives and assessment needs to evolve to the point where a second and third activity monitoring device were added to the study as the first device was phased out, making inclusion of early and late joining subjects extremely difficult. We assess the performance of a rank-based approach for estimating the latent correlation to transform activity counts from various devices to a common scale along with approaches using quantile and regression based transformations and quantify the amount of shared information across devices.

email: jordan.johns@jhu.edu

56l. IMPROVED DOUBLY ROBUST ESTIMATION IN LEARNING INDIVIDUALIZED TREATMENT RULES

Yinghao Pan*, University of North Carolina, Charlotte
Yingqi Zhao, Fred Hutchinson Cancer Research Center

Due to patient's heterogeneous response to treatment, there is a growing interest in developing novel and efficient statistical methods in estimating individualized treatment rules (ITRs). The central idea is to recommend treatment according to patient characteristics, and the optimal ITR is the one that maximizes the expected clinical outcome if followed by the patient population. We propose an improved estimator of the optimal ITR that enjoys two key properties. First, it is doubly robust, meaning that the proposed estimator is consistent if either the propensity score or the outcome model is correct. Second, it achieves the smallest variance among its class of doubly robust estimators when the propensity score model is correctly specified, regardless of the specification of the outcome model. Simulation studies show that the estimated optimal ITR obtained from our method yields better clinical outcome than its main competitors. Data from Sequenced Treatment Alternatives to Relieve Depression (STAR*D) study is analyzed as an illustrative example.

email: ypan8@uncc.edu

56m. PROJECTING CLINICAL TRIAL RESULTS TO ALTERNATIVE POPULATIONS BY INTERPOLATION

Shuang Li*, Southern Methodist University
Daniel F. Heitjan, Southern Methodist University and University of Texas Southwestern Medical Center

The population in which one tests a treatment can differ from the population in which one intends to apply it. We propose a method to project causal estimates from a clinical trial to reflect the distributions of covariates in the target population. We assume that trial patients are similar to population patients, other than on a set of measured factors whose distribution in the population is known. We start with computing estimates of treatment effects in strata defined by these factors, and reweighting the estimates to match the distribution in the target population. Interpolation, as we call the method, can reduce bias in cases where there is a treatment-by-covariate interaction and the distribution of the covariate in the target population differs from that in the trial. We apply the idea to the standard intention-to-treat estimate and to the instrumental variable estimate of the complier-average causal effect, describing adjustment for both discrete and continuous stratification factors. We demonstrate our method in two examples --- the Lipids Research Clinics Coronary Primary Prevention Trial and the New York School Choice Experiment.

email: shuangli@smu.edu

57. AGREEMENT MEASURES AND DIAGNOSTICS

THE IMPACT OF RATER FACTORS ON ORDINAL AGREEMENT

Kerrie Nelson*, Boston University
Aya Mitani, Boston University
Don Edwards, University of South Carolina

Ordinal categorical scales are commonly used in screening and diagnostic tests to classify a patient's disease status such as mammography. However, discrepancies are often observed between experts' classifications. We present a flexible model-based approach based upon the class of generalized linear mixed models to assess the impact of rater characteristics including experience and training on agreement and association between many raters' ordinal classifications. We demonstrate our proposed approach in a recent large-scale breast cancer study.

email: kerrie@bu.edu

ROBUST MATRIX-BASED MEASURES OF AGREEMENT BASED ON L-STATISTICS FOR REPEATED MEASURES

Elahe Tashakor*, Pennsylvania State College of Medicine
Vernon M. Chinchilli, Pennsylvania State College of Medicine

The concordance correlation coefficient is commonly used to assess agreement between two raters or two methods of measuring a response when the data are measured on a continuous scale. It typically is used under the assumption that data are normally distributed. However, in many practical applications, data often are skewed and/or thick-tailed. Previously, we have proposed an approach that extends existing methods of robust estimators to produce more robust versions of the concordance correlation coefficient. In this paper we extend the application of this class of estimators to a multivariate situation, possibly repeated measurements, based on a matrix norm that possesses the properties needed to characterize the level of agreement between two $p \times 1$ vectors of random variables. We provide two data examples to illustrate the methodology, and we discuss the results of computer simulation studies that evaluate statistical performance.

email: eqt5124@psu.edu

NET BENEFIT OF A DIAGNOSTIC TEST TO RULE-IN OR RULE-OUT A CLINICAL CONDITION

Gene A. Pennello*, U.S.Food and Drug Administration

Some diagnostic tests are intended to rule-in or rule-out a clinical condition. For example, in pregnant women suspected to be at risk for pre-term delivery, fFN assays are used rule out significant risk. Net benefit of a diagnostic test is the difference in expected utility between the test and a random test. Net-benefit of ruling out (in) is the difference in expected utility between the test and a random test that is always positive (negative) because in this setting the standard of care is to consider subjects as having (not having) the condition. Net benefit measures can be used to optimize test cut-offs and set performance goals. Net benefits of a test relative to a perfect test can be displayed in a likelihood ratio graph. To illustrate, I use measures of diagnostic accuracy for predictive markers (Simon, JNCI, 2015) to show that in one study the relative net benefit of EGFR status is higher for ruling out than ruling in treatment of non-small cell lung cancer patients with gefitinib.

email: gene.pennello@fda.hhs.gov

ESTIMATION OF SENSITIVITY AND SPECIFICITY OF MULTIPLE TESTS AND DISEASE PREVALENCE FOR REPEATED MEASURES WITHOUT GOLD STANDARD

Chunling Wang*, University of South Carolina
Timothy E. Hanson, Medtronic Inc.

A novel latent class model for multiple dependent diagnostic tests applied repeatedly over time is proposed for the case where there is no gold standard. The model is identifiable with as few as three tests; this is in contrast to non-repeated measures where the untenable assumption of conditional independence must be made in order to obtain inference. Pairwise covariance parameters among tests at each time point in the diseased and non-diseased populations are explicitly included, as well as temporal covariance parameters. An efficient componentwise adaptive Markov chain Monte Carlo scheme that requires no tuning is developed for posterior inference. The proposed model is broadly illustrated via simulations and an analysis of repeated measures scaphoid fracture data, for which there is no gold standard.

email: chunling@email.sc.edu

A NEW MEASURE OF DIAGNOSTIC ACCURACY WITH CUT-POINT SELECTION CRITERION FOR K-STAGE DISEASES USING CONCORDANCE AND DISCORDANCE

Jing Kersey*, Georgia Southern University
Hani Samawi, Georgia Southern University
Jingjing Yin, Georgia Southern University
Haresh Rochani, Georgia Southern University
Xianyan Zhang, Georgia Southern University
Chen Mo, Georgia Southern University

An essential aspect for medical diagnostic testing using biomarkers is to find an optimal cut-point which categorizes a patient as diseased or healthy. This aspect can be extended to the diseases which can be classified into more than two classes. For diseases with general k (k>2) stages, well-established measures include hypervolume under the manifold and the generalized Youden Index. Another two diagnostic accuracy measures, maximum absolute determinant (MADET) and Kullback-Leibler divergence measure (KL), are recently proposed. This research proposes a new measure of diagnostic accuracy based on concordance and discordance (CD) for diseases with k (k>2) stages and uses it as a cut-points selection criterion. The CD measure utilizes all the classification information and provides more balanced class probabilities in some scenarios. Power studies and simulations will be carried out to compare the performance of available measures. As well, an example of an actual dataset from the medical field will be provided using the proposed CD measure.

email: jingkersey@gmail.com

58. VARIABLE SELECTION

VARIABLE SELECTION IN ENRICHED DIRICHLET PROCESS WITH APPLICATIONS TO CAUSAL INFERENCE

Kumaresh Dhara*, University of Florida
Michael J. Daniels, University of Florida

Dirichlet process mixtures are often used to model the joint distribution of a response and predictors. However, the clusters formed when fitting the model often depends heavily on the covariates. Enriched Dirichlet process priors (EDP) overcomes these issues by modeling the joint distribution of response and predictors using a nested structure. EDP has been recently used in causal inference. It is common that a large number of covariates are available for modeling the response but only a few of them are important. In this paper, we propose a variable selection approach while using an enriched Dirichlet process. Removing irrelevant covariates helps in efficient and simpler modeling of the joint structure of the response and covariates.

email: k.dhara@ufl.edu

MODEL CONFIDENCE BOUNDS FOR VARIABLE SELECTION

Yang Li*, Renmin University of China

In this article, we introduce the concept of model confidence bounds (MCBs) for variable selection in the context of nested models. Similarly to the endpoints in the familiar confidence interval for parameter estimation, the MCBs identify two nested models (upper and lower confidence bound models) containing the true model at a given level of confidence. Instead of trusting a single selected model obtained from a given model selection method, the MCBs proposes a group of nested models as candidates and the MCBs' width and composition enable the practitioner to assess the overall model selection uncertainty. A new graphical tool — the model uncertainty curve (MUC) — is introduced to visualize the variability of model selection and to compare different model selection procedures. The MCBs methodology is implemented by a fast bootstrap algorithm that is shown to yield the correct asymptotic coverage under rather general conditions. Our Monte Carlo simulations and a real data example confirm the validity and illustrate the advantages of the proposed method.

email: yang.li@ruc.edu.cn

ALL MODELS ARE WRONG BUT MANY ARE USEFUL: VARIABLE IMPORTANCE FOR BLACK-BOX, PROPRIETARY, OR UNKNOWN PREDICTION MODELS, WITH MODEL CLASS RELIANCE

Aaron J. Fisher*, Harvard T. H. Chan School of Public Health
Cynthia Rudin, Duke University
Francesca Dominici, Harvard University

Variable importance (VI) tools describe how much covariates contribute to a prediction model's accuracy. However, important variables for one well-performing model may be unimportant for another model. We propose Model Class Reliance (MCR) as the range of VI values across all well-performing model in a prespecified class (e.g. all linear models of dimension p). Thus, MCR accounts for the fact that many prediction models may fit the data well. In the process of deriving MCR, we show several informative results for permutation-based VI estimates, similar to the VI measures used in Random Forests. Specifically, we derive connections between permutation importance estimates for a single prediction model, U-statistics, conditional causal effects, and linear model coefficients. We then give probabilistic bounds for MCR, using a novel technique that can also be applied to attain finite-sample bounds for many other problems. We apply MCR in a public dataset of Broward County criminal records to study the reliance of recidivism prediction models on sex and race. In this application, MCR can be used to help inform VI for unknown, proprietary models.

email: aafisher@hsph.harvard.edu

VARIABLE SCREENING WITH MULTIPLE STUDIES

Tianzhou Ma*, University of Maryland, College Park
Zhao Ren, University of Pittsburgh
George Tseng, University of Pittsburgh

Advancement in technology has generated abundant high-dimensional data that allows integration of multiple relevant studies. Due to huge computational advantage, variable screening methods based on marginal correlation have become promising alternatives to the popular regularization methods for variable selection. However, all screening methods are limited to single study so far. We consider a general framework for variable screening with multiple related studies, and further propose a novel two-step screening procedure using a self-normalized estimator for high-dimensional regression analysis in this framework. Compared to the one-step and rank-based procedures, our procedure greatly reduces false negative errors while keeping a low false positive rate. Theoretically, we show that our procedure possesses the sure screening property with weaker assumptions on signal strengths and allows the number of features to grow at an exponential rate of the sample size. Simulations and a real transcriptomic application illustrate the advantage of our method. Extension of the current method to model-free and distributionally robust statistics for general use will also be discussed.

email: tma0929@umd.edu

SIMULTANEOUS ESTIMATION AND VARIABLE SELECTION FOR INTERVAL-CENSORED DATA WITH BROKEN ADAPTIVE RIDGE REGRESSION

Qiwei Wu*, University of Missouri, Columbia
Hui Zhao, Zhongnan University of Economics and Law, China
Gang Li, University of California, Los Angeles
Jianguo Sun, University of Missouri, Columbia

The simultaneous estimation and variable selection for Cox model has been discussed by several authors (Fan and Li, 2002; Huang and Ma, 2010; Tibshirani, 1997) when one observes right-censored failure time data. However, there does not seem to exist an established procedure for interval-censored data, a more general and complex type of failure time data, except two parametric procedures in Scolas et al. (2016) and Wu and Cook (2015). To address this, we propose a broken adaptive ridge (BAR) regression procedure that combines the strengths of the quadratic regularization and the adaptive weighted bridge shrinkage. In particular, the method allows for the number of covariates to be diverging with the sample size. Under some weak regularity conditions, unlike most of the existing variable selection methods, we establish both the oracle property and the grouping effect of the proposed BAR procedure. We conduct an extensive simulation study and show that the proposed approach works well in practical situations and deals with the collinearity problem better than the other oracle-like methods. An application is also provided.

email: qw235@mail.missouri.edu

AN APPLICATION OF PENALIZED QUASI-LIKELIHOOD IN VARIABLE SELECTION ON PARAMETRIC ACCELERATED FAILURE TIME MODELS WITH FRAILTY

Sarbesh R. Pandeya*, Georgia Southern University
Lili Yu, Georgia Southern University
Hani M. Samawi, Georgia Southern University
Xinyan Zhang, Georgia Southern University

Variable selection has been administered in many studies of Biostatistics especially regarding large multi-center clinical trials that have high dimensional data points. The penalized likelihood is one of the most prominent methods for variable selection as it has shown a consistent and successful range of essential and necessary variables. Therefore, we aim to comparatively test these different penalty functions to conduct

variable selection among parametric accelerated failure time (AFT) models with mixed effects (i.e., with a shared frailty parameter). The determination of the best estimates was done via mean sums of squares and the computational costs. We propose to use the penalized quasi-likelihood (PQL) approach with an induced penalty to our selection process. We used this method to compare with other penalty functions in mixed models and evaluated their performance under censoring.

email: sp03459@georgiasouthern.edu

SIMULTANEOUS SELECTION AND INFERENCE FOR VARYING COEFFICIENTS WITH ZERO REGIONS: A SOFT-THRESHOLDING APPROACH

Yuan Yang*, University of Michigan
Jian Kang, University of Michigan
Yi Li, University of Michigan

Varying coefficient models have emerged as an important tool to explore dynamic patterns in many scientific areas, such as biomedicine, economics, finance, politics, and epidemiology. An often overlooked aspect, however, is that some varying coefficients may have regions where the effects are zero. For example, in a preoperative opioid use study that motivates this paper, it was found that no association between opioid use and pain level exists among patients whose BMI is larger than 30 or less than 25, while a dose-response relationship exists among those with BMIs between 25 and 30. Most existing methods focus on estimation and variable selection, ignoring detection of zero regions. Therefore, we propose a new soft-thresholded varying coefficient model that enables us to perform variable selection and detect the zero regions of selected variables simultaneously and to obtain point estimates of the varying coefficients with zero regions and construct the associated sparse confidence intervals. We show that the proposed method enjoys good theoretical properties and achieves the desired coverage probability.

email: yuanyang@umich.edu

59. CAUSAL INFERENCE

CAUSAL ISOTONIC REGRESSION

Ted Westling*, University of Pennsylvania
Peter Gilbert, Fred Hutchinson Cancer Research Center
Marco Carone, University of Washington

In observational studies, potential confounders may distort the causal relationship between an exposure and an outcome. However, under some conditions, a causal dose-response curve can be recovered via the G-formula. Most classical methods for estimating such curves rely on restrictive parametric assumptions, which carry risk of model misspecification. Nonparametric estimation in this context is challenging because many available nonparametric estimators are sensitive to the selection of certain tuning parameters, and performing valid inference with such estimators can be difficult. In this work, we propose a nonparametric estimator of a causal dose-response curve known to be monotone. We show that our proposed estimation procedure generalizes the classical least-squares isotonic regression estimator of a monotone regression function. We describe theoretical properties of our proposed estimator, including its irregular limit distribution and the potential for doubly-robust inference. Furthermore, we illustrate its performance via numerical studies, and use it to assess the effect of BMI on immune response in HIV vaccine trials.

email: tgwest@pennmedicine.upenn.edu

MULTIPLY ROBUST TWO-SAMPLE INSTRUMENTAL VARIABLE ESTIMATION

BaoLuo Sun*, National University of Singapore

Although instrumental variable (IV) methods are widely used to estimate causal effects in the presence of unmeasured confounding, the IVs, exposure and outcome are often not measured in the same sample due to complex data harvesting procedures. Following the influential articles by Klevmarken (1982) and Angrist & Krueger (1992, 1995), numerous empirical researchers have applied two-sample IV methods to perform joint estimation based on an IV-exposure sample and an IV-outcome sample. We develop a general semi-parametric framework for two-sample data combination models from a missing data perspective, and characterizes the efficiency bound based on the full data model. In the context of the two-sample IV problem as a specific example, the framework offers insights on issues of efficiency and robustness of existing estimators. We propose new multiply robust locally efficient estimators of the causal effect of exposure on the outcome, and illustrate the methods through simulation and an econometric application on public housing projects.

email: stasb@nus.edu.sg

AN INSTRUMENTAL VARIABLE FOR COX MODELS EXTENDED TO NON-PROPORTIONAL HAZARDS AND EFFECT MODIFICATION

James O'Malley*, Geisel School of Medicine at Dartmouth
Pablo Martinez-Camblor, Geisel School of Medicine at Dartmouth
Todd MacKenzie, Geisel School of Medicine at Dartmouth
Douglas O. Staiger, Dartmouth College
Philip P. Goodney, Geisel School of Medicine at Dartmouth
Two-stage instrumental variable methods are commonly used for estimating average causal effects in the presence of an unmeasured confounder. In the context of the proportional hazard Cox regression models, this problem has recently received attention with several methods being proposed. We developed an improved estimator under the incumbent two stage residual inclusion (2SRI) procedure by adding a Gaussian frailty in the second stage and apply it to the situation in which both the treatment and the unmeasured confounder effects can be time-varying or the treatment effect is modified by an observed covariate. We show that when the effect of the unmeasured confounder and/or the treatment change during the follow-up, the first stage of the 2SRI algorithm induces a frailty with time-varying coefficients in the second stage. A Monte Carlo simulation study demonstrates the superior performance of the proposed extension of 2SRI we develop. We apply the new procedure to estimate the effect of endarterectomy (versus carotid artery stenting) on the time to death of patients suffering from carotid artery disease using linked Vascular Quality Initiative Registry – Medicare data.

email: James.OMalley@Dartmouth.edu

POST-RANDOMIZATION BIOMARKER EFFECT MODIFICATION IN AN HIV VACCINE CLINICAL TRIAL

Bryan S. Blette*, University of North Carolina, Chapel Hill
Peter B. Gilbert, University of Washington
Bryan E. Shepherd, Vanderbilt University
Michael G. Hudgens, University of North Carolina, Chapel Hill

The most recent HIV vaccine trial (HVTN 505) showed no overall efficacy of the tested vaccine to prevent HIV infection. However, several immune response markers were strongly correlated with infection in vaccine recipients, suggesting that a qualitative interaction may have occurred. Current principal stratification effect modification (PSEM) methods make untestable structural infection risk assumptions and more assumption-lean PSEM methods are needed to assess a qualitative interaction hypothesis. Notably, many methods from the survivor average causal effect (SACE) literature rely on leaner assumption sets; we show that these can be adapted to the PSEM problem in the special case of a binary intermediate response variable and map this adaptation. This opens up a host of new PSEM methods for a binary intermediate variable measured via two-phase sampling, for a dichotomous or failure time final outcome, and including or excluding the SACE monotonicity assumption. The new methods support that the vaccine partially protected recipients with a high polyfunctional CD8+ T cell response, which may support further research into vaccines designed to improve this type of response.

email: blette@live.unc.edu

PRINCIPAL STRATIFICATION FOR CAUSAL EFFECTS CONDITIONING ON A BINARY POST-TREATMENT VARIABLE IN CLINICAL TRIALS

Judah Abberbock*, GlaxoSmithKline
Gong Tang, University of Pittsburgh

Early-stage breast cancer trials are often conducted as neoadjuvant trials where therapy is administered prior to surgical removal of breast tumors. In these trials, pathological complete response (pCR) at time of surgery is an intercurrent event as defined in the ICH E9(R1) Addendum. We proposed a method under the principal stratification framework to estimate the treatment effect in a long-term binary outcome among those who would achieve pCR if given the new treatment. With a baseline auxiliary variable, we imposed a logistic regression model for predicting the counterfactual intercurrent outcome pCR given treatment, baseline characteristics and the observed long-term outcome. Under a monotonicity assumption that pCR responders on the control arm would respond to treatment, the regression parameters are identifiable and estimated via the general method of moments. Subsequently an imputation procedure is adopted to estimate the treatment effect for the principal strata of interest. We compared the performance of our proposed method with other approaches in simulation studies. Data from a neoadjuvant breast cancer clinical trial are used to demonstrate the proposed method.

email: judah.x.abberbock@gsk.com

60. GENETIC EFFECTS/HERITABILITY

PHENOTYPE IMPUTATION INTEGRATING GWAS SUMMARY ASSOCIATION STATISTICS, DEEP PHENOTYPED COHORTS AND LARGE BIOBANKS: APPLICATION TO NICOTINE UPTAKE PHENOTYPES

Lina Yang*, Pennsylvania State College of Medicine
Dajiang Liu, Pennsylvania State College of Medicine

Phenotype imputation predicts unmeasured traits using estimated trait correlations. Existing methods such as MICE or SOFTIMPUTE rely solely on trait correlation, but ignore genetic data. A new method PHENIX integrates genetic data and explicitly models genetic correlation, but fails to integrate GWAS summary statistics from large datasets and is computationally intensive. We developed a novel phenotype imputation method based on Gaussian graphic models, which effectively and efficiently integrates GWAS summary statistics, biobanks with genetic but less detailed phenotypic data, and datasets with rich phenotypes. The new method greatly improves the estimates of genetic correlation and hence the imputation accuracy. Applying phenotype imputation, we imputed cotinine levels - a smoking biomarker, into UK Biobank (N=148505) based upon the detailed phenotypes from the PASS smoking trial and GSCAN consortium GWAS summary statistics on common smoking traits. We identified 2 novel loci, which improved the power of an earlier study using only measured cotinine levels (N=4548). We expect our approach to be a valuable tool to integrate small deep-phenotyped studies with large biobanks.

email: lzy51@psu.edu

A UNIFIED METHOD FOR RARE VARIANT ANALYSIS OF GENE-ENVIRONMENT INTERACTIONS

Elise Lim*, Boston University
Han Chen, University of Texas Health Science Center at Houston
José Dupuis, Boston University
Ching-Ti Liu, Boston University

Advanced technology in whole-genome sequencing has offered the opportunity to investigate the genetic contribution, particularly rare variants, to complex traits. Many rare variants analysis methods have been developed to jointly model the marginal effect but gene-environment (GE) interactions are understudied. We develop a unified method to detect GE interactions of a set of rare variants using generalized linear mixed effects model. The proposed method can accommodate both binary and continuous traits in related or unrelated samples. We implement a variance component score test to reduce the computational burden. Our simulation study shows that the proposed method maintains correct type I error rates and high power under various scenarios. We illustrate our method to test gene x smoking interaction on body mass index in the Framingham Heart Study and replicate the CHRNB4 gene association reported in previous consortium meta-analysis of single nucleotide polymorphism-smoking interaction. Our proposed GE test is efficient and is applicable to both binary and continuous phenotypes, while appropriately accounting for familial or cryptic relatedness.

email: elise625@gmail.com

HERITABILITY ESTIMATION AND GENETIC ASSOCIATION TESTING IN LONGITUDINAL TWIN STUDIES

Souvik Seal*, University of Minnesota
Saonli Basu, University of Minnesota

A longitudinal twin study involves repeated measurements on twins and can serve as a powerful resource to detect genetic association in the development of complex traits. Extending any familial dependency structure in the context of a longitudinal study is immensely challenging because it requires handling additional two modes of dependence; the dependence over time within an individual and the dependence over time between any pair of relatives in a family. In this paper, we investigate the challenges in modeling a longitudinal twin study in light of heritability estimation and genetic association testing. In such a setup, the most common technique for association testing, is to use a linear mixed model with a correlation structure which is kronecker product of two matrices corresponding to those two modes of dependence, whereas for heritability estimation, a multivariate version of the classical twin ACE model is used. Our model which connects these two schools of thought neatly with more flexible assumptions and is an extension of the traditional Falconer's approach. We also propose two rapid estimation procedures of the model parameters.

email: sealx017@umn.edu

COMPARISON OF HYPOTHESIS TESTING METHODS ON RANDOM GENETIC EFFECTS IN FAMILY DATA

Nicholas DeVogel*, Medical College of Wisconsin
Tao Wang, Medical College of Wisconsin

Adjusting for genetic similarity between individuals is an important aspect to consider when analyzing family data. This similarity can come from two sources, additive and dominance, that can be incorporated as random effects in a linear mixed model to induce the genetic correlation between the individuals. However, testing whether additive or dominance effects exist, or are necessary to adjust for, leads to a non-standard hypothesis test on the variance components that places the components on the boundary of the parameter space. Additionally problematic is the existence of nuisance variance components. As such, no standard testing method exists for this situation. This study compares different variance component testing procedures for this non-standard test on additive or dominance familial genetic effects in linear mixed models. The testing procedures include restricted likelihood ratio, F-statistic, and score tests. The comparison of these tests will be based on false positive rates and powers estimated from simulations studies where family structured data is simulated from the 1000 Genomes Project with varying levels of true genetic correlation.

email: ndevogel@mcw.edu

SMALL AND LARGE SAMPLE BIAS OF REML ESTIMATES OF GENOMIC HERITABILITY ESTIMATES: AN ASSESSMENT USING BIG DATA

Raka Mandal*, Michigan State University
Tapabrata Maiti, Michigan State University
Gustavo De Los Campos, Michigan State University

In genetics, the trait heritability represents the proportion of variance of a phenotype that can be explained by genetic factors. Recently, there has been an increased interest on the estimation of genomic heritability, that is the proportion of variance of a trait or in disease risk that can be explained by regression on large sets of molecular markers (e.g., SNPs).The debate about the methodology has been largely based on results from simulation studies which can produce, depending on the simulation settings, from nearly unbiased to seriously biased estimators. The recent availability of very large biomedical datasets present numerous opportunities for assessing the sampling properties of REML estimates. In this study we use real data from UK-Biobank (N~100K, K=1000) to investigate the effects of sample size and model complexity (number of SNPs, from 5K to ~600K) on estimates of genomic heritability using human height as an example trait. We use recursive partitioning of the training data and show that the average estimator of the genomic heritability decreases with sample size; we conclude that the popular REML estimates of genomic heritability can be seriously biased.

email: mandalr1@msu.edu

AN ADAPTIVE TEST FOR HIGH-DIMENSIONAL GENERALIZED LINEAR MODELS WITH APPLICATION TO DETECT GENE-ENVIRONMENT INTERACTIONS

Chong Wu*, Florida State University
Gongjun Xu, University of Michigan
Xiaotong Shen, University of Minnesota
Wei Pan, University of Minnesota

In spite of its urgent importance in the era of big data, testing high-dimensional parameters in generalized linear models (GLMs) in the presence of high-dimensional nuisance parameters has been largely under-studied, especially with regard to constructing powerful tests for general (and unknown) alternatives. Most existing tests are powerful only against certain alternatives and may yield incorrect Type 1 error rates under high-dimensional nuisance parameter situations. In this paper, we propose the adaptive interaction sum of powered score (aiSPU) test in the framework of penalized regression with a non-convex penalty, called truncated Lasso penalty (TLP), which can maintain correct Type 1 error rates while yielding high statistical power across a wide range of alternatives. To calculate its p-values analytically, we derive its asymptotic null distribution. Via simulations, its superior finite-sample performance is demonstrated over several representative existing methods. In addition, we apply it and other representative tests to an Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, detecting possible gene-gender interactions for Alzheimer's disease.

email: cwu3@fsu.edu

61. COMPUTATIONAL METHODS AND MASSIVE DATA SETS

ON THE USE OF OPTIMAL TRANSPORTATION THEORY TO MERGE DATABASES: APPLICATION TO CLINICAL RESEARCH

Nicolas J. Savy*, Toulouse Institute of Mathematics
Valérie Garès, INSA of Rennes
Chloé Dimeglio, CHU of Toulouse
Gregory Guernec, INSERM unit 1027
Benoit Lepage, INSERM unit 1027
Michael R. Kosorok, University of North Carolina, Chapel Hill
Philippe Saint-Pierre, Toulouse Institute of Mathematics

We consider the problem of finding a relevant way to merge two databases when a variable of interest is not coded on the same scale in both databases. To address this issue, an algorithm, based on the optimal transportation, is proposed if an individual identification common to the two databases is not possible. Optimal transportation theory gives us an application to transport the measure associated with the variable in database A to the measure associated with the same variable in database B. To do so, a cost function has to be introduced and an allocation rule has to be defined. Such a function and such a rule is proposed involving the information contained in the covariates. Our method is compared to multiple imputation by chained equation and has demonstrated a better performance in many situations. Applications on both simulated and real datasets show that the efficiency of the proposed merging algorithm depends on how the covariates are linked with the variable of interest.

email: Nicolas.Savy@math.univ-toulouse.fr

AN ONLINE UPDATING APPROACH FOR TESTING THE PROPORTIONAL HAZARDS ASSUMPTION WITH STREAMS OF BIG SURVIVAL DATA

Yishu Xue*, University of Connecticut
HaiYing Wang, University of Connecticut
Jun Yan, University of Connecticut
Elizabeth D. Schifano, University of Connecticut

The Cox model, which remains as the first choice in analyzing time-to-event data even for large datasets, relies on the proportional hazards (PH) assumption. When the data size exceeds the computer memory, the standard statistics for testing the PH assumption can no longer be easily calculated. We propose an online updating approach with minimal storage requirement that updates the standard test statistic as each new block of data becomes available. Under the null hypothesis of PH, the proposed statistic is shown to have the same asymptotic distribution as the standard version if it could be computed with a super computer. In simulation studies, the test and its variant based on most recent data blocks maintain their sizes when the PH assumption holds and have substantial power to detect different violations of the PH assumption. The approach is illustrated with the survival analysis of patients with lymphoma cancer from the Surveillance, Epidemiology, and End Results Program. The proposed test promptly identified deviation from the PH assumption that was not captured by the test based on the entire data.

email: yishu.xue@uconn.edu

REAL-TIME REGRESSION ANALYSIS OF STREAMING HEALTH DATASETS

Lan Luo*, University of Michigan
Peter X.K. Song, University of Michigan

Perpetual data collection takes place in many health science areas such as national disease registry, mobile health, and disease surveillance, where large volumes of observations become sequentially available over time. This paper is motivated by an analysis of kidney transplant data that are continually updated over time by the Scientific Registry of Transplant Recipients (SRTR) of the United States. We develop a fast real-time approach for estimation and inference, in which both parameter estimates and their standard errors are updated along with data updates. In the implementation, we propose an incremental Newton-Raphson algorithm in the Lambda architecture with Spark to facilitate real-time data analysis. Both estimation consistency and asymptotic normality of our proposed estimator are established, and the Wald test is utilized to conduct real-time inference. We illustrate performance of our method with simulation experiments and analysis of the SRTR data, and find that both donor's and recipient's age, as well as type of organ donor, are among the most significant variables associated with the risk of five-year graft failure.

email: luolsph@umich.edu

APPLICATION OF DEEP CONVOLUTIONAL NEURAL NETWORKS IN CLASSIFICATION OF PROTEIN SUBCELLULAR LOCALIZATION WITH MICROSCOPY IMAGES

Mengli Xiao*, University of Minnesota
Xiaotong Shen, University of Minnesota
Wei Pan, University of Minnesota

Single cell microscopy images analysis has proved invaluable in protein subcellular localization for inferring gene/protein function. Fluorescent-tagged proteins across cellular compartments are tracked and imaged in response to genetic or environmental perturbations. With a large amount of images generated by high-content microscopy while manual labeling is both labor-intensive and error-prone, machine learning offers a viable alternative for automatic labeling of subcellular localizations. On the other hand, applications of deep learning methods to large datasets in natural images have become quite successful. An appeal of deep learning methods is that they can learn salient features from complicated data with little data preprocessing. For such purposes, we applied several representative and state-of-the-art architectures of deep Convolutional Neural Networks (CNNs) and two popular ensemble methods, random forests and gradient boosting, to predict protein subcellular localization with a cell image dataset. We show the consistently better predictive performance of CNNs over the two ensemble methods. We also demonstrate the use of CNNs for feature extraction.

email: xiaox345@umn.edu

EXTRACTING THE COMMON PATTERN BETWEEN HIGH-DIMENSIONAL DATASETS

Zhe Qu*, Tulane University
Mac Hyman, Tulane University

One popular model for jointly analyzing two high-dimensional data sets on a common set of objects is to decompose each of the two possibly unequal-sized data matrices into three parts: a low-rank "common" matrix, a low-rank "distinctive" matrix, and an additive noise matrix. Existing decomposition methods claim that their common matrices have captured the common pattern of the two data sets. However, their so-called common pattern only refers to the common latent factors and fails to take into account the weights (a.k.a. coefficients) of these latent factors. We develop a novel method for extracting both the common latent factors and their common weights to form the common pattern as their matrix product. The proposed method is applied to analyze the common pattern of the left-hand and the right-hand activations in human brain using the motor-task functional MRI data from the Human Connectome Project.

email: zqu2@tulane.edu

PROJECTION INFERENCE FOR PENALIZED REGRESSION ESTIMATORS

Biyue Dai*, University of Iowa
Patrick Breheny, University of Iowa

In recent years, many efforts have been made in the field of high-dimensional inference. We propose an intuitive and computationally efficient procedure to carry out inference for high-dimensional MCP-penalized linear models. The KKT conditions for MCP allow inference to be carried out using an approximate projection onto the column space of the active features. We show that this approach, PIPE (Projection Inference using Penalized regression Estimators), can be used to construct confidence intervals and false discovery rates for MCP estimates. We conducted simulations to study PIPE's empirical performance at finite sample size and compare the approach to existing high-dimensional inference methods. Finally, we consider the potential of extending this idea to estimates arising from penalized regression models using the LASSO.

email: biyue-dai@uiowa.edu

62. RECENT ADVANCES IN BAYESIAN NETWORK META-ANALYSIS

BAYESIAN NETWORK META-REGRESSION FOR ORDINAL OUTCOMES: APPLICATIONS TO COMPARING CROHN'S DISEASE TREATMENTS

Joseph G. Ibrahim*, University of North Carolina, Chapel Hill
Yeongjin Gwon, University of Connecticut
Ming-Hui Chen, University of Connecticut
May Mo, Amgen Inc.
Tony Jiang, Amgen Inc.
Amy Xia, Amgen Inc.

Logistic regression models are frequently used to model ordinal response data due to the attractive proportional odds property. In this talk, we propose a new network-meta regression approach for modeling ordinal outcomes under different links. Specifically, we develop regression model based on aggregate treatment-level covariates for the underlying cut-off points of the ordinal outcomes as well as aggregate trial-level covariates for the variances of the random effects to capture heterogeneity across trials. We also examine the importance of links in fitting ordinal responses in the middle categories. This novel theoretical development allows us to incorporate a variety of links regardless of symmetry or asymmetry. A case study demonstrating the usefulness of the proposed methodology is carried out using aggregate ordinal outcome data from 17 clinical trials for treating Crohn's disease.

email: ibrahim@bios.unc.edu

BAYESIAN JOINT NETWORK META-REGRESSION METHODS ADJUSTING FOR POST-RANDOMIZATION VARIABLES

Jing Zhang*, University of Maryland
Mark Wymer, University of Maryland
Qinshu Lian, Genentech
Haitao Chu, University of Minnesota

Though there is board consensus on accounting for study-level covariates in mixed treatment comparisons through network meta-regression, it is much more challenging to adjust for postrandomization variables, which are expected to differ between treatment arms within a study. Examples include differential noncompliance, measured as the proportion of premature treatment discontinuation or drop out, loss to follow-up, or change to an alternative therapy. In existing network meta-regression methods, study-level covariates are assumed to be fixed. However, postrandomization variables are generally considered random and thus cannot be adjusted for by existing methods. In this paper (talk), we will propose novel Bayesian joint network meta-regression methods to account for post-randomization variables, which enables more accurate estimation of treatment effects. We will illustrate the proposed methods through simulations and real data analyses.

email: jzhang86@umd.edu

BAYESIAN INCONSISTENCY DETECTION FOR NETWORK META-ANALYSIS

Ming-Hui Chen*, University of Connecticut
Hao Li, University of Connecticut
Cheng Zhang, University of Connecticut
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Arvind K. Shah, Merck & Co.
Jianxin Lin, Merck & Co.

Many clinical trials have been carried out on safety and efficacy evaluation of cholesterol lowering drugs. To synthesize the results from different clinical trials, we examine treatment level (aggregate) network meta-data from 29 double-blind, randomized, active or placebo-controlled clinical trials with statins or statins plus Ezetimibe on adult treatment-naive patients with primary hypercholesterolemia. In this paper, we construct general linear hypotheses to investigate consistency under a general fixed effects model without any assumptions. Some interesting results are established on equivalence between consistency assumptions on the treatment effect parameters and the hypotheses about certain contrasts of

parameters. A general algorithm is developed to compute the contrast matrix under consistency assumptions. Furthermore, a new Bayesian approach is developed to detect inconsistency. Simulation studies are carried out. We apply the proposed methodology to conduct an analysis of the network meta-data from 29 trials with 11 treatment arms.

email: ming-hui.chen@uconn.edu

BAYESIAN HIERARCHICAL MODELS FOR N-OF-1 TRIALS WITH ORDINAL OUTCOMES

Youdan Wang*, Brown University
Christopher Schmid, Brown University

N-of-1 trials are single-patient multiple crossover experiments in which patients switch between two or more treatments. Although N-of-1 trials are designed to estimate treatment efficacy in single patients, N-of-1 trials assessing the same scientific questions may be combined together and analyzed with a multilevel model. When the treatments compared differ between trials, it may be possible to construct a network of treatments and use network meta-analysis methods to make comparisons among the treatments both at the individual and population levels. Ordinal outcomes are very common in clinical trials as well as in social science research, but methods published for network meta-analysis of ordinal data are very limited. We develop Bayesian network analytical models for combining N-of-1 trials with ordinal outcomes and will compare results with those from analyses of individual data alone. We will use data from a series of N-of-1 trials assessing different treatments for chronic pain to demonstrate the application of hierarchical models for N-of-1 ordinal data.

email: youdan_wang@yahoo.com

63. STATISTICAL METHODS TO SUPPORT VALID AND EFFICIENT USE OF ELECTRONIC HEALTH RECORDS DATA

ENABLING IMPRECISE EHR DATA FOR PRECISION MEDICINE

Tianxi Cai*, Harvard University

While traditional cohort studies and clinical trials remain critical sources for studying disease risk, progression and treatment response, they have limitations including the generalizability of the study findings to the real world and the limited ability to test broader hypotheses. In recent years, large electronic health records (EHR) data integrated with biological data now exist as a new source for precision medicine research. These datasets open new opportunities for deriving real-word, data-driven prediction models of disease risk and progression. Yet, they also bring methodological challenges. For example, obtaining precise clinical event onset time, is a major bottleneck in EHR research, as it requires laborious medical record review and such information may not be accurately documented. In this talk, I'll discuss statistical methods for developing risk prediction models that can efficiently leverage both a small partially labeled dataset and a large unlabeled data. These methods will be illustrated using EHR data from Partner's Healthcare.

email: tcai@hsph.harvard.edu

A BAYESIAN NONPARAMETRICS FOR MISSING DATA IN EHRs

Michael J. Daniels*, University of Florida

We propose a Bayesian nonparametric approach to address selection bias for inference using electronic health records (EHRs). Data provenance, the collection of decisions that give rise to the observed data, is modularized and modeled to properly adjust for the selection bias using mixture models. The approach is used to assess the long terms effects of bariatric surgery.

email: mdaniels@stat.ufl.edu

SAMPLING DESIGNS FOR RESOURCE EFFICIENT COLLECTION OF OUTCOME LABELS FOR STATISTICAL LEARNING, WITH APPLICATIONS TO ELECTRONIC MEDICAL RECORDS

Patrick J. Heagerty*, University of Washington
Wei Ling Katherine Tan, University of Washington

In leveraging data derived from large-scale electronic medical record (EMR) systems for research, an important first step is the accurate identification of key clinical outcomes. Some outcomes are recorded in structured data, while other outcomes must be derived or predicted from both structured and unstructured data. Statistical classification of clinical outcomes requires the collection of a training set, which is a sample where actual binary outcomes are abstracted and labeled by human medical experts. When the outcome is rare, simple random sampling for abstraction results in very few cases. Such outcome class imbalance results in insufficient information for classifier development, yet additional abstraction is often expensive and time-consuming. In this work, we propose sampling designs for outcome label collection and subsequent statistical learning targeting the rare outcome scenario. Our proposed designs are amenable for valid analysis, and are more resource efficient, requiring a smaller sample size for modeling goals compared to conventional simple random sampling.

email: heagerty@uw.edu

ACCOUNTING FOR DIFFERENTIAL MISCLASSIFICATION IN EHR-DERIVED PHENOTYPES

Rebecca A. Hubbard*, University of Pennsylvania

Many electronic health records (EHR)-derived phenotypes are imperfect and suffer from exposure-dependent differential misclassification due to variability in the quality and availability of EHR data across exposure groups. For instance, a breast cancer recurrence phenotype may be more sensitive and less specific in women with comorbidities because they are in more frequent contact with the healthcare system than healthier women. This leads to differential outcome misclassification in subsequent analyses. Through simulations and analyses of real EHR-derived data, we demonstrate the effect of differential misclassification on bias and type I error of analyses using imperfect phenotypes.We then present alternative approaches to accounting for differential error in phenotypes including bias reduction approaches that incorporate information on uncertainty derived from statistical phenotyping approaches. Because differential outcome misclassification, which is expected to

be particularly common for EHR-derived outcomes, induces bias and may lead to spurious findings, we conclude with recommendations for best practices to improve the validity of research using EHR-derived outcomes.

email: rhubb@pennmedicine.upenn.edu

64. RECENT ADVANCES IN THE ANALYSIS OF TIME-TO-EVENT OUTCOMES SUBJECT TO A TERMINAL EVENT

FLEXIBLE ACCELERATED TIME MODELING OF RECURRENT EVENTS DATA IN THE PRESENCE OF A DEPENDENT TERMINAL EVENT

Limin Peng*, Emory University
Bo Wei, Emory University
Zhumin Zhang, University of Wisconsin, Madison
HuiChuan Lai, University of Wisconsin, Madison

Accelerated time modeling provides a useful alternative prospective for assessing covariate events on recurrent event outcomes, which renders physical interpretations. The generalized accelerated recurrence time model (GART) significantly extends the traditional accelerated failure time model for recurrent events, offering extra flexibility in accommodating heterogeneous covariate effects. In practice, the observation of recurrent events is often stopped by a dependent terminal event. To address such a realistic scenario, we discuss two extensions of the GART models that can appropriately account for the presence of a dependent terminal event. We develop estimation and inference procedures for both extensions of the GART model, and establish desirable asymptotic properties. The proposed estimation and inference procedures can be readily implemented based on existing software. Simulation studies demonstrate good finite-sample performance of the proposed methods. We illustrate the proposed methods via an application to a dataset from the Cystic Fibrosis Foundation Patient Registry (CFFPR).

email: lpeng@sph.emory.edu

BAYESIAN VARIABLE SELECTION FOR A SEMI-COMPETING RISKS MODEL WITH THREE HAZARD FUNCTIONS

Andrew G. Chapple*, Louisiana State University School of Public Health
Marina Vannucci, Rice University
Peter F. Thall, University of Texas MD Anderson Cancer Center
Steven Lin, University of Texas MD Anderson Cancer Center

A variable selection procedure is developed for a semi-competing risks regression model with three hazard functions that uses spike-and-slab priors and stochastic search variable selection algorithms for posterior inference. A rule is devised for choosing the threshold on the marginal posterior probability of variable inclusion based on the Deviance Information Criterion (DIC) that is examined in a simulation study. The method is applied to data from esophageal cancer patients from the MD Anderson Cancer Center, Houston, TX, where the most important covariates are selected in each of the hazards of effusion, death before effusion, and death after effusion. The DIC procedure that is proposed leads to similar selected models regardless of the choices of some of the hyperparameters. The application results show that patients with intensity-modulated radiation therapy have significantly reduced risks of pericardial effusion, pleural effusion, and death before either effusion type.

email: achapp@lsuhsc.edu

A GENERALIZED NESTED CASE-CONTROL DESIGN FOR THE SEMI-COMPETING RISKS SETTING

Ina Jazic*, Vertex Pharmaceuticals
Tianxi Cai, Harvard T.H. Chan School of Public Health
Sebastien Haneuse, Harvard T.H. Chan School of Public Health

When certain covariates of interest are difficult to obtain, researchers may designate a subsample of patients on whom to collect complete data: one way is using the nested case-control (NCC) design, in which risk set sampling is performed based on a single outcome. Recent work has extended the NCC study design to accommodate the simultaneous analysis of multiple outcomes, including the semi-competing risks setting, where some non-terminal event is subject to a terminal event. Estimation and inference for a weighted illness-death model have been proposed for this setting, as well as the supplemented nested case-control study (SNCC) design, allowing an initial NCC study to be supplemented with cases of the non-index outcome. For more common outcomes, however, selecting all cases into an NCC study may not be feasible. We propose a generalized NCC design that allows for subsamples of both index and non-index cases to be selected, with appropriate inverse probability weights. We investigate operating characteristics and design considerations via simulation, and we illustrate our methods on a study of acute graft-versus-host disease after hematopoietic stem cell transplantation.

email: Ina_Jazic@vrtx.com

ANALYSIS OF SEMI-COMPETING RISKS DATA VIA BIVARIATE LONGITUDINAL MODELS

Daniel Nevo*, Tel Aviv University
Sebastien Haneuse, Harvard T.H. Chan School of Public Health

An example of semi-competing risk data analysis is the study of Alzheimer's disease and death. Most existing methods to evaluate risk factors treat the dependence between Alzheimer's and death occurrence as nuisance and restrict it to follow simple mathematical model. However, these methods may suffer from misspecification of the dependence structure. Furthermore, information about the dependence, including its form, trajectory over time and how it depends on covariates can provide new clinical knowledge. Therefore, we propose a new framework for analyzing semi-competing risks data by the means of bivariate longitudinal modeling. Our methods differentiate between local and global dependence. Local dependence captures the co-occurrence of Alzheimer's and deaths within a short period of time, while global dependence is the long-term effect of Alzheimer's on the risk of death. We incorporate flexible splines into our models to account for changes over time and develop a penalized maximum likelihood estimators and associated inference for the parameters of interest.

email: danielnevo@gmail.com

65. RECENT ADVANCES IN STATISTICAL METHODS FOR PRECISION MEDICINE

VARIABLE SELECTION IN JOINT FRAILTY MODELS OF RECURRENT AND TERMINAL EVENTS

Lei Liu*, Washington University in St. Louis
Dongxiao Han, Chinese Academy of Sciences
Liuquan Sun, Chinese Academy of Sciences
Xiaogang Su, University of Texas at El Paso
Zhou Zhang, Northwestern University Feinberg School of Medicine

Recurrent event data are commonly encountered in biomedical studies. In many situations, they are subject to an informative terminal event, e.g., death. Joint modeling recurrent and terminal events has attracted substantial recent research interest. On the other hand, there may exist a large number of covariates in such data. How to conduct variable selection for joint frailty proportional hazards models has become a challenge in practical data analysis. We tackle this issue on the basis of the "Minimum approximated Information Criterion" (MIC) method. The proposed method can be conveniently implemented in SAS Proc NLMIXED for both normal and log-Gamma frailties. The finite sample behavior of the proposed estimators is evaluated through simulation studies. The proposed method is applied to model recurrent opportunistic diseases in the presence of death in an AIDS study.

email: lei.liu@wustl.edu

SUBGROUP IDENTIFICATION USING ELECTRONIC HEALTH RECORD DATA

Marianthi Markatou*, State University of New York at Buffalo

In this talk, we first present a systematic review of methods developed to identify subgroups or cohorts of patients that exhibit specific phenotypes using Electronic Health Records (EHR) data. The review draws heavily from the biomedical informatics literature, and in particular, from the work performed by various established scientific networks. We discuss the many challenges associated with the identification of these subgroups using EHR data, and evaluate the performance of these algorithms using a variety of metrics. We then discuss a method that appears promising for increasing the accuracy of identifying patients with the phenotype of interest.

email: markatou@buffalo.edu

HUMAN DISEASE NETWORK (HDN) ANALYSIS OF DISEASE PREVALENCE

Shuangge Ma*, Yale University

In "classic" biomedical research, the prevalence of different diseases has been separately. Accumulating evidences have suggested that diseases can be "correlated". In most existing studies, such correlation has only been studied for a small number of pre-selected diseases. In our study, we conduct big-data analysis of the Taiwan National Health Insurance Research Database (NHIRD). Novel HDN analysis is conducted to "globally" examine the interconnections among disease prevalence. Important characteristics of the HDN are quantified.

email: shuangge.ma@yale.edu

MODELING TIME-VARYING EFFECTS OF MULTILEVEL RISK FACTORS OF HOSPITALIZATIONS IN PATIENTS ON DIALYSIS

Damla Senturk*, University of California, Los Angeles
Yihao Li, University of California, Los Angeles
Danh V. Nguyen, University of California, Irvine
Yanjun Chen, University of California, Irvine
Connie M. Rhee, University of California, Irvine
Kamyar Kalantar-Zadeh, University of California, Irvine

For chronic dialysis patients, a unique population requiring continuous medical care, methodologies to monitor patient outcomes, such as hospitalizations, over time, after initiation of dialysis, are of particular interest. Contributing to patient hospitalizations are a number of multilevel covariates such as demographics and comorbidities at the patient-level, and staffing composition at the dialysis facility-level. We propose a varying coefficient model for multilevel risk factors (VCM-MR) to study the time-varying effects of covariates on patient hospitalization risk as a function of time on dialysis. The proposed VCM-MR also includes subject-specific random effects to account for within-subject correlation and dialysis facility-specific fixed effect varying coefficient functions. The method is applied to model hospitalization risk using the rich hierarchical data available on dialysis patients initiating dialysis between January 1, 2006 and December 31, 2008 from United States Renal Data System, a large national database, where 331,443 hospitalizations over time are nested within patients, and 89,889 patients are nested within 2,201 dialysis facilities.

email: dsenturk@ucla.edu

66. CHALLENGES AND ADVANCES IN WEARABLE TECHNOLOGY

CLASSIFICATION OF HUMAN ACTIVITY BASED ON THE RAW ACCELEROMETRY DATA: COMPARISON OF THE DATA TRANSFORMATIONS

Jaroslaw Harezlak*, Indiana University, Bloomington
Marcin Straczkiewicz, Indiana University, Bloomington
Jacek Urbanek, Johns Hopkins University

Objective characterization of person's physical activity is an important task in medicine and public health. The daily level of person's activity is highly correlated with human health and it is often used as a marker of physical fitness. To ensure objective physical activity summaries, scientists utilize accelerometers and sophisticated computational algorithms to analyze the raw data collected by them. In our work, we compare the algorithms using a variety of raw data transformations to classify the human activity as walking, standing and sitting/lying. Specifically, we use the spherical coordinate system in the time domain, short time Fourier transform and continuous wavelet transform. We show that depending on the activity performed there is not a universal "best" method and utilization of a methods' ensemble is beneficial to classify and summarize the aspects of human activity.

email: harezlak@iu.edu

SUB-SECOND LEVEL ACCELEROMETRY DATA IN HEALTH RESEARCH: CHALLENGES AND OPPORTUNITIES

Jacek K. Urbanek*, Johns Hopkins University School of Medicine
Marta Karas, Johns Hopkins Bloomberg School of Public Health
Marcin Straczkiewicz, Indiana University, Bloomington
Jaroslaw Harezlak, Indiana University, Bloomington
Jiawei Bai, Johns Hopkins Bloomberg School of Public Health
Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health
Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Wearable accelerometers provide detailed, objective, and continuous measurements of physical activity (PA). Recent advances in technology and the decreasing cost of wearable devices led to an explosion in the popularity of wearable technology in health research. An ever increasing number of studies collect high-throughput, sub-second level raw accelerometry data. In this presentation we discuss problems related to the collection and analysis of such data and provide insights into potential solutions. In particular, we describe the size and complexity of the data, the within- and between-subject variability and the effects of sensor location on the body. We illustrate these points using the Developmental Epidemiological Cohort Study (DECOS), which collected raw accelerometry data on individuals both in a controlled and the free-living environment.

email: jurbane2@jhu.edu

PHYSICAL ACTIVITY VERSUS INACTIVITY VERSUS SLEEP: NONPARAMETRIC ESTIMATES OF ISOTEMPORAL SUBSTITUTION EFFECTS

John W. Staudenmayer*, University of Massachusetts Amherst

It has long been recognized that how people allocate their time between physical activity, sedentary behavior, and sleep influences health. Large representative epidemiological surveillance studies such as the National Health and Nutrition Examination Survey (NHANES) recently have added data to that discussion by both assessing health outcomes and using accelerometers to estimate how people spend their time at different levels of physical activity, inactivity, and sleep. Those measurements and estimates present a modeling challenge since the physical activity, inactivity, and sleep covariates add up to a constant (24 hours) for each participant. An ordinary linear regression model approach to this type of data has received a lot of attention in the epidemiology literature. When $p-1$ of the covariates are used in a model a regression coefficient estimates the effect of increasing its covariate by one unit on the health outcome while decreasing the left out covariate by one unit; i.e. an isotemporal substitution effect. This talk will develop a novel non-parametric approach to this problem.

email: jstauden@me.com

DERIVING OBJECTIVE ACTIVITY MEASURES FROM ACCELEROMETRY DATA FOR MORTALITY PREDICTION MODELS

Ekaterina Smirnova*, Virginia Commonwealth University
Andrew Leroux, Johns Hopkins Bloomberg School of Public Health
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Systematic assessment of population mortality risks and evaluation of behavioral patterns associated with mortality is critical for public health and preventive care. A growing number of studies use accelerometers to accurately access population physical activity in free-living conditions. The raw data is typically summarized into a minute-level accelerometry count measure, which leads to functional data collected over 1440 minutes per each subject day. In this presentation, we describe the data collection challenges (missing measurements, extensive device non-wear time) and the activity summary level (minute-by-minute data, total activity counts, active to sedentary and sedentary to active transitional probability, functional principal components scores) that reflect meaningful clinical information related to mortality risks assessment. We illustrate these challenges on the example of building the 5-year all-cause mortality prediction model using National Health Examination Study (NHANES) data.

email: ekaterina.smirnova@vcuhealth.org

67. STATISTICAL MODELING IN CELL BIOLOGY

MISSING DATA AND TECHNICAL VARIABILITY IN SINGLE-CELL RNA-SEQUENCING EXPERIMENTS

Stephanie Hicks*, Johns Hopkins Bloomberg School of Public Health
William Townes, Harvard T.H. Chan School of Public Health
Martin Aryee, Massachusetts General Hospital
Rafael Irizarry, Dana-Farber Cancer Institute

Single-cell RNA-Seq (scRNA-seq) is the most widely used high-throughput technology to measure genome-wide gene expression at the single-cell level. Unlike bulk RNA-Seq, the majority of reported expression levels in scRNA-seq are zeros and the proportion of genes reporting the expression level to be zero varies substantially across cells. However, it remains unclear to what extent this cell-to-cell variation is being driven by technical versus biological variation. Here, we use an assessment experiment to examine data from published studies. We present evidence that some of these zeros are driven by technical variation by demonstrating that scRNA-seq produces more zeros than expected and that this bias is greater for lower expressed genes. This missing data problem is exacerbated by the fact that technical variation varies cell-to-cell, which can be confused with novel biological results. Finally, we propose a cell-specific censoring with a varying-censoring aware matrix factorization model (VAMF) for dimensionality reduction that permits the identification of factors in the presence of the above described systematic bias.

email: shicks19@jhu.edu

FITTING STOCHASTIC MODELS TO IN VIVO CELL LINEAGE TRACKING DATA

Jason Xu*, Duke University
Samson Koelle, University of Washington
Peter Guttorp, University of Washington
Chuanfeng Wu, National Heart, Lung, and Blood Institute, National Institutes of Health
Cynthia Dunbar, National Heart, Lung, and Blood Institute, National Institutes of Health
Janis Abkowitwz, University of Washington
Vladimir Minin, University of California, Irvine

Recent cell lineage tracking techniques are producing significant insights toward better understanding complex biological systems. However, the development of statistical methodology for analyzing the resulting data is still in its early stages. We present a class of multi-type branching processes to model hematopoiesis, the process of blood cell production, together with a moment estimator to fit time series data from cell barcoding experiments of primate hematopoiesis. The proposed estimator is computationally efficient and accounts for missing data inherent to the experimental design. In contrast to prior statistical modeling efforts focusing on stem cell behavior, our framework ascribes parameters corresponding to the fate decisions of intermediate progenitor cells, and estimates these parameters jointly. This further allows us to compare potential model structures through cross-validation toward answering competing hypotheses about the differentiation pathways in such systems. The methodology is broadly transferrable to related models arising in systems biology, epidemiology, and other fields.

email: jason.q.xu@duke.edu

IMPUTATION OF SINGLE-CELL GENE EXPRESSION WITH AUTOENCODER NEURAL NETWORKS

Audrey Q. Fu*, University of Idaho

Single-cell RNA-sequencing (scRNA-seq) is a rapidly evolving technology that allows users to measure gene expression levels at an unprecedented resolution. Despite the explosive growth in the number of cells that can be assayed by a single experiment, scRNA-seq data continue to suffer from technical limitations, including high rates of dropouts. Dropout events result in a large number of genes with zero count, which complicate downstream analyses of scRNA_seq data. To overcome this problem, we developed deep learning algorithms LATE and TRANSLATE to impute scRNA-seq data. LATE (Learning with AuToEncoder) trains an autoencoder de novo using scRNA-seq data to correct the count estimates for lowly expressed genes, and TRANSLATE (TRANSfer learning with LATE) allows the use of a reference gene expression panel to provide LATE with an initial set of parameter estimates. On both simulated and real data, LATE and TRANSLATE outperform existing scRNA-seq imputation methods. Importantly, LATE and TRANSLATE are highly scalable and can impute gene expression levels in over 1 million cells in just a few hours.

email: audreyf@uidaho.edu

MODELS FOR DEPENDENT DATA IN SINGLE CELL ASSAYS

Andrew McDavid*, University of Rochester
Corey Kimzey, University of Rochester

Gene expression profiling of single cells (scRNAseq) and flow cytometry have refined and defined new cell types and states. Focus is increasingly turning towards population-based studies that rely on nested designs in which a cohort of individuals is repeatedly measured by sampling their cells. It has also been observed that even putatively independent designs will generate dependent data when batch effects are present. I consider a two-part, zero-inflated log-Normal random effects model for these dependent data. The variance parameters may be weakly identified in a given transcript, but related across the various transcripts measured. To leverage this property, I propose a hierarchical model that borrows information on the variance parameters across markers.

email: Andrew_McDavid@urmc.rochester.edu

68. DIAGNOSTICS, ROC, AND RISK PREDICTION

ON KULLBACK-LEIBLER DIVERGENCE AS A MEASURE FOR MEDICAL DIAGNOSTICS AND CUT-POINT SELECTION CRITERION

Hani Samawi*, Georgia Southern University
Jingjing Yin, Georgia Southern University
Xinyan Zhang, Georgia Southern University
Lili Yu, Georgia Southern University
Haresh Rochani, Georgia Southern University
Robert Vogel, Georgia Southern University

Recently, Kullback-Leibler divergence measur (KL), which captures the disparity between two distributions, has been considered as an index for determining the diagnostic performance of markers. This study investigates variety of applications of KL divergence in medical diagnostics, including overall measures of rule-in and rule-out potential and proposes an optimization criteria based on KL divergence for cut point selection. Moreover, the paper links the KL divergence with some common Receiver Operating Characteristic (ROC) measures and presents analytically and numerically the relations in situations of one crossing point as well as multiple crossing points. Furthermore, the graphical application and interpretation of KL divergence, which is referred as the information graph, is discussed. A comprehensive data analysis of the Dutch Breast Cancer Data are provided to illustrate the proposed applications.

email: samawi.hani2@gmail.com

RECEIVER OPERATING CHARACTERISTIC CURVES AND CONFIDENCE BANDS FOR SUPPORT VECTOR MACHINES

Daniel J. Luckett*, University of North Carolina, Chapel Hill
Eric B. Laber, North Carolina State University
Michael R. Kosorok, University of North Carolina, Chapel Hill

Many problems that appear in biomedical decision making, such as diagnosing disease and predicting response to treatment, can be expressed as binary classification problems. The costs of false positives and false negatives vary across application domains and receiver operating characteristic (ROC) curves provide a visual representation of this trade-off. Nonparametric estimators for the ROC curve, such as a weighted support vector machine (SVM), are desirable because they are robust to model misspecification. While weighted SVMs have great potential for estimating ROC curves, their theoretical properties were heretofore underdeveloped. We propose a method for constructing confidence bands for the SVM ROC curve and provide the theoretical justification for the SVM ROC curve by showing that the risk function of the estimated decision rule is uniformly consistent across the weight parameter. We demonstrate the proposed confidence band method and the superior sensitivity and specificity of the weighted SVM compared to commonly used methods using simulation studies. We present two illustrative examples: diagnosis of hepatitis C and predicting response to cancer treatment.

email: luckett@live.unc.edu

PROPER ROC MODELS: DUAL BETA MODEL AND WEIGHTED POWER FUNCTION MODEL

Hongying Peng*, University of Cincinnati
Douglas Mossman, University of Cincinnati
Marepalli Rao, University of Cincinnati

In biomedical studies, receiver operating characteristic (ROC) curve is a standard tool for assessing diagnostic accuracy at distinguishing two mutually exclusive conditions (e.g., “disease” vs. “no disease”). The conventional binormal ROC curve-fitting method usually produces improper ROC curves with a convex portion leading to misinformation for diagnostic accuracy. We propose and evaluate two novel, simple and flexible two-parameter ROC curve fitting methods, named the Weighted Power Function (WPF) model and Dual Beta (DB) model, which always produce proper ROC curves with desired properties such as monotonically decreasing slope and simple formulas for the area under the curve. The DB and WPF models produce smaller Root Mean Square Errors than the binormal model even when samples are simulated based on the binormal model assumption. This suggests that DB and WPF models offer accurate and flexible modeling strategies for ROC curves.

email: penghn@mail.uc.edu

A GENERAL FRAMEWORK FOR USING THE OVERALL CONCORDANCE STATISTIC TO ASSESS THE DISCRIMINATORY ABILITY OF RISK PREDICTIONS

Li C. Cheung*, National Cancer Institute, National Institutes of Health
Qing Pan, The George Washington University
Barry Graubard, National Cancer Institute, National Institutes of Health

Harrell’s concordance (C) index is a widely used non-parametric measure of the discrimination ability of a risk prediction or biomarker. However, the value of the C-index depends on study-specific characteristics, such as the length of follow-up, the censoring distribution, and the population representativeness of the data. In addition to bias issues, some commonly used software provide overly conservative variance estimates. Building upon previously published work, we propose a modified concordance index that measures overall discriminatory ability up to a time tau and that converges to a censoring-independent quantity. This modified C-index can be applied to interval- or right-censored event time data or to data that arise from complex survey designs, allowing for more meaningful comparison of concordance statistics estimated from different data sets. We derive closed-form variance for the modified C-index using Taylor linearization methods that can be implemented using available software such as the R survey package. Results from our simulation studies suggest that the modified C-index and proposed variance methods will perform well in finite samples.

email: li.cheung@nih.gov

EVALUATING PREDICTIVE ACCURACY OF TRADITIONAL AND MACHINE LEARNING BASED SURVIVAL MODELS

Yue-Ming Chen*, Merck Research Laboratories
Dai Feng, Merck Research Laboratories
Nicholas C. Henderson, Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University
Vladimir Svetnik, Merck Research Laboratories

In this study we aim at evaluating predictive accuracy of some traditional and machine learning based survival models that can be used for both low and high dimensional data. The prediction models considered include the traditional regression modeling strategies such as semiparametric Cox regression and parametric accelerated failure time (AFT) model, and state of the art machine learning approaches such as random survival forests and Bayesian additive regression trees (BART) based survival models. For BART, we study a method that represents the nonparametric likelihood for the Kaplan–Meier (KM) estimator in a form suitable for original BART, and another fully nonparametric AFT model using BART to model the mean effect and centered Dirichlet process to model the residual. We assess and compare predictions using different measures including C-statistics and AUC. We use a variety of simulated and real-world data to evaluate the performance of different methods.

email: yue-ming.chen@merck.com

69. MICROBIOME DATA: FINDING ASSOCIATIONS AND TESTING

ROBUST SCREENING FOR ASSOCIATIONS BETWEEN MICROBIOME COMMUNITY PROFILES AND A LARGE NUMBER OF INDIVIDUAL GENOMIC OUTCOMES

Weijia Fu*, Fred Hutchinson Cancer Research Center
Nanxun Ma, Fred Hutchinson Cancer Research Center
Michael C. Wu, Fred Hutchinson Cancer Research Center

Assessing associations between microbiome composition and other omics markers is increasingly popular as such information provides clues as to key biological mechanisms. Although a wide range of methods have been developed for associating microbiome composition with individual outcomes, such methods may not be appropriate when screening a large number of markers as genomic markers are subject to irregular distributions with outliers and heteroskedasticity. Thus, we propose to use a robust kernel machine test to rapid screen for associations between individual genomic outcomes and overall microbiome composition. The approach is based on robust regression which allows for irregularities in the genomic outcome distribution. Use of the kernel framework allows for embedding of important ecological and phylogenetic information for the microbiome data. Asymptotic null distribution is derived but finite sample approximations are also presented. Simulations and real data analyses show that the proposed methods improve both type I error control and power.

email: fuwj0528@uw.edu

IMPROVED VARIANCE COMPONENT SCORE TESTS OF MARKER-ENVIRONMENT INTERACTIONS

Nanxun Ma*, University of Washington
Michael C. Wu, Fred Hutchinson Cancer Research Center
Jing Ma, Fred Hutchinson Cancer Research Center

Variance component score test is popular for assessing interactions between large number of markers and environmental exposure. Existing methods tend to give inflated type I error due to challenges in estimating the null model which contains main effects. Least squares based estimation fails when the number of markers increases, and ridge regression estimates for the main effects gives biased estimates due to penalty. We propose an improved variance component score test that estimates the main effects under the null hypothesis using bias-corrected ridge regression. We adapt the usual variance component test using a novel empirical corrected projection matrices corresponding to the bias-corrected ridge regression to construct test statistic. Asymptotically, VC statistic follows a mixture of chi-squares distribution, but we develop a small sample approach to accurately estimate p-values even when the same size is modest. Simulations and real data analysis demonstrate that the bias-corrected interaction tests improve type I error control compared with current methods, while maintaining power. We apply the approach to a microbiome study and identify microbiome composition interactions.

email: nanxunma@uw.edu

ECOLOGICAL DISSIMILARITIES FOR PAIRED AND LONGITUDINAL MICROBIOME ASSOCIATION ANALYSIS

Anna M. Plantinga*, Williams College
Michael C. Wu, Fred Hutchinson Cancer Research Center

Due to the substantial inter-subject variability in the human microbiome, paired and longitudinal studies of the microbiome have become increasingly popular as a way to reduce unmeasured confounding and increase statistical power. However, few methods are available for analyzing such datasets. We introduce a paired UniFrac dissimilarity that summarizes within-individual changes in microbiome composition and then compares these compositional shifts across individuals. This dissimilarity depends on a novel transformation of relative abundances, which we then extend to more than two time points and incorporate into several phylogenetic and non-phylogenetic dissimilarities. The data transformations and resulting dissimilarities may be used in a wide variety of downstream analyses, including ordination analysis and distance-based hypothesis testing. Simulations demonstrate that tests based on these dissimilarities have proper type 1 error and high power. We apply the method in two real datasets and compare the results to single time point analyses.

email: amp9@williams.edu

AN ADAPTIVE DISTANCE-BASED KERNEL ASSOCIATION TEST BASED ON THE GENERALIZED LINEAR MIXED EFFECT MODEL FOR CORRELATED MICROBIOME STUDIES

Hyunwook Koh*, Johns Hopkins Bloomberg School of Public Health
Ni Zhao, Johns Hopkins Bloomberg School of Public Health

For correlated (e.g., family-based or longitudinal) microbiome studies, a distance-based kernel association test based on the linear mixed effect model, namely, cSKAT, has recently been introduced. cSKAT models the microbial community based on an ecological (e.g., UniFrac) distance measure and tests its association with a host phenotype (e.g., health/disease status). The use of ecological distance measures renders a higher power to cSKAT than the ones based on non-ecological distance measures. However, cSKAT is limited to handle a normally distributed phenotype and the item-by-item use of the distance measures. The power performance of cSKAT differs a lot by which distance measure is in use, yet it is also highly challenging to choose an optimal distance measure to use because of the unknown nature of the true association. Here, we introduce a distance-based kernel association test based on the generalized mixed effect model, namely, GLMM-MiRKAT, to handle any exponential family-based phenotype. Notably, for a robust power performance not depending on the choice of distance measure, we propose a data-driven adaptive test of GLMM-MiRKAT, namely, aGLMM-MiRKAT.

email: hkoh7@jhu.edu

AN ADAPTIVE MULTIVARIATE TWO-SAMPLE TEST WITH APPLICATION TO MICROBIOME DIFFERENTIAL ABUNDANCE ANALYSIS

Xiang Zhan*, The Pennsylvania State University

There are two different modes of microbiome differential abundance analyses (MDAA): the individual-based univariate MDAA and the group-based multivariate

MDAA. The univariate analysis identifies differentially abundant taxa under certain statistical error measurements such as FDR, which is typically complicated by the high-dimensionality of taxa and complex correlation structure among taxa. The multivariate analysis evaluates the overall shift in the abundance of microbiome composition between two conditions, which provides useful information for the necessity of follow-up univariate analysis. In this paper, we present a novel adaptive multivariate two-sample test for the MDAA to examine whether the composition of a group of taxa are different between two conditions. Our simulation studies and real data applications demonstrated that the new test was often more powerful than several competing methods while preserving the correct type I error rate.

email: xiangzhan9@gmail.com

SEMIPARAMETRIC METHODS FOR TESTING FOR DIFFERENTIAL ABUNDANCE IN MICROBIOME STUDIES

Olivier Thas*, Hasselt University, Belgium, Ghent University, Belgium and University of Wollongong, Australia
Leyla Kodalci, Hasselt University, Belgium
Stijn Hawinkel, Ghent University, Belgium

Microbiome count data can be considered as compositional mutivariate observations, which are characterised by a sum-constraint. Many data analysis methods developed for compositional data make use of log-ratios of counts, but these are problematic in the presence of many zero counts, as is the case for microbiome. In this talk we focus on testing for differential abundance. We have developed several semiparametric methods for compositional microbiome data. The methods do not rely on strong distributional assumptions, avoid log-ratios and account for library size variability. The methods make use of either means, rank or sign statistics. False discovery rate control happens through a new permutation method that accounts for the discreteness of the p-value null distributions. Results from a realistic simulation study suggest that the new FDR-method perfoms well and that particularly the sign-based methods perform well for overdispersed microbiome data.

email: Olivier.Thas@UHasselt.be

70. COMPARATIVE EFFECTIVENESS, CLUSTERED AND CATEGORICAL DATA

A POTENTIAL OUTCOMES APPROACH TO SUBGROUP DISCOVERY IN COST-EFFECTIVENESS ANALYSIS

Nicholas A. Illenberger*, University of Pennsylvania
Nandita Mitra, University of Pennsylvania
Andrew J. Spieker, Vanderbilt University Medical Center

Informed health policy decisions require consideration of both treatment cost and clinical efficacy. In recent work, we introduced the Cost-Effectiveness Determination (CED) curve as a clinically interpretable alternative to existing measures. This approach uses the potential outcomes framework to elucidate the proportion of individuals experiencing greater overall treatment benefit as a function of the willingness-to-pay. We develop methodology to estimate covariate-specific CED curves to account for effect modification. Estimation and inference for conditional CED curves can be performed using generalized linear models, together with Monte-Carlo integration to correct for confounders. Methods that allow us to determine which patient subgroups experience greater treatment benefit promotes efficient use of resources and tailored healthcare decisions. As an illustration, we use SEER-Medicare data to compare radiation therapy to no treatment in post-surgery endometrial cancer patients.

email: nillen@pennmedicine.upenn.edu

ESTIMATING TREATMENT IMPORTANCE IN MULTIDRUG-RESISTANT TUBERCULOSIS USING TARGETED LEARNING: AN OBSERVATIONAL INDIVIDUAL PATIENT DATA META-ANALYSIS

Guanbo Wang*, McGill University

Multi-drug-resistant tuberculosis (MDR-TB) is defined as strains of TB that do not respond to at least the two most powerful anti-TB drugs. It is always treated with multiple antibiotics. Our data consist of individual patient data from 31 international observational studies. In this study, we develop identifiability criteria for the estimation of a generalized treatment importance metric in the context where not all drugs are observed in all studies. With stronger causal assumptions, this treatment importance can be interpreted as the effect of adding a medication to the existing regimen. We then use this metric to rank 15 observed drugs in terms of their estimated add-on value. Using the concept of transportability, we propose an implementation of targeted maximum likelihood estimation (TMLE), a doubly robust and locally efficient plug-in estimator, to estimate the treatment importance metric. A clustered sandwich estimator is adopted to compute variance estimates and produce confidence intervals. Simulation studies are conducted to assess the performance of our estimator, verify the double robustness property, and assess the appropriateness of the variance estimation approach.

email: guanbo.wang@mail.mcgill.ca

LIMITED INFORMATION EMPIRICAL BAYES FOR CLASSIFICATION OF SUBJECTS ACROSS CONDITIONS

Hillary Koch*, The Pennsylvania State University
Siqi Xiang, University of North Carolina, Chapel Hill
Han Wang, Zhejiang University
Feipeng Zhang, Hunan Normal University
Qunhua Li, The Pennsylvania State University

Genomics studies often seek to uncover relationships among many subjects (e.g. gene expression levels) across many conditions (e.g. tissues or cell types). The simplest approach to identifying significant relationships across conditions first identifies relationships of interest condition-by-condition, then uses these individual outcomes to determine shared relationships across conditions. This technique does not leverage any information regarding the dependence among conditions, and is thus severely underpowered. Several advancements over condition-by-condition analyses have been developed, yet each still suffers from some critical drawbacks both in model flexibility and computational intractability for even a modest number of conditions. Using a novel limited information mixture model, we sidestep computational intractibility by paring down the collection of classes each subject may belong to. We then feed this output into an empirical Bayesian framework

which simultaneously performs classification and parameter estimation without making restrictive assumptions on correlations between and sizes of the effects (e.g. expression levels) of interest.

email: hbk5086@psu.edu

CONDITIONAL LOG-LINEAR REGRESSION FOR STRATIFIED PAIRS OF ORDINAL RESPONSES

Jing Yu*, University of North Carolina, Chapel Hill
Gary Koch, University of North Carolina, Chapel Hill

Stratified matching enables elimination of variability in the outcome among the matched sets, thus making it useful for designing a clinical study. In each of 79 clinics for a skin condition treatment study, one patient received the treatment, and another patient received a placebo. Variables collected included age, sex, and an initial grade, which ranged from 1 to 4 for mild to severe. The data are collected for each patient in each clinic, including the assignments, the skin condition improvement (as 2, 1 or 0) and the baseline variables. The primary objective of this clinical trial is to demonstrate whether the test treatment provides better improvement than the placebo, after controlling for baseline covariates. We develop methodology to estimate the treatment effect using the equal adjacent odds ratio model with adjustment for covariates. We also can further address extensions of covariance adjustment for ordinal outcomes through using both non-parametric strategies and logistic regression methods. The non-parametric methods have essentially no assumptions for a randomized clinical trial, and for a log-linear model they produce results with expected properties.

email: yujingwy@gmail.com

LATENT CLASS MODEL FOR FINDING CO-OCCURRENT PATTERNS IN PROCESS DATA

Guanhua Fang*, Columbia University
Zhiliang Ying, Columbia University
Jingchen Liu, Columbia University

Process data, temporally ordered data with categorical observations, are of great interest for researchers due to its massive underlying information. Each process is a collection of multi-type events along with time stamps, recording how an individual performs or reacts in a given time period. As opposed to traditional cross-sectional response data, process data entails much more features and patterns which could be useful for interpretation of human characteristics. Therefore, it is calling for new effective exploratory analysis tools. We introduce a latent theme dictionary model (LTDM) for modeling event processes to identify co-occurrent event patterns as well as to cluster individuals with similar behaviors into sub-groups. We propose a non-parametric Bayes LTDM algorithm by using a Markov chain Monte Carlo method for model estimation. The algorithm performs well on the simulated data sets confirming the theoretical results. We also apply our method to log files from the "Traffic Item" in the 2012 Programme for International Student Assessment and obtain interesting findings.

email: gf2340@columbia.edu

ACCURACY OF LATENT CLASS ITEM RESPONSE THEORY MODELS FOR EXAMINING MEASUREMENT INVARIANCE IN PATIENT-REPORTED OUTCOMES MEASURES

Tolulope Sajobi*, University of Calgary
Richard Sawatzky, Trinity Western University
Lara Russell, University of British Columbia
Oluwaseyi A. Lawal, University of Calgary
Juxin Liu, University of Saskatchewan
Bruno D. Zumbo, University of British Columbia
Lisa M. Lix, University of Manitoba

Latent class item response theory (IRT) models have recently been proposed for examining measurement invariance in patient-reported outcome (PRO) instruments. However, there is limited investigation of the accuracy of these mixture models under a variety of data analytic conditions. This study compares the accuracy of conventional IRT and latent class IRT models for detecting measurement invariance (MI) in ordinal PRO data. Monte Carlo methods were used to evaluate the performance of latent class 2-parameter graded response model in detecting MI in ordinal items under a variety of data analytic conditions. Simulation conditions investigated included number of latent classes, number of items, distribution of latent factor across classes, sample size, and levels of item responses. Bias and standard error of item parameters, misclassification error, and latent class mean differences were used to assess the accuracy of these models. Data from the Alberta Provincinal Project on Outcome Assessment in Coronary Heart Disease (APPROACH), a provincial population-based registry of cardiac patients, was used to demonstrate the implementation of these methods.

email: ttsajobi@ucalgary.ca

71. CAUSAL INFERENCE AND MEASUREMENT ERROR

INSTRUMENTAL VARIABLE APPROACH TO ESTIMATING THE SCALAR-ON-FUNCTION REGRESSION MODEL WITH MEASUREMENT ERROR WITH APPLICATION TO ENERGY EXPENDITURE ASSESSMENT IN CHILDHOOD OBESITY

Carmen D. Tekwe*, Texas A&M University
Roger S. Zoh, Texas A&M University
Lan Xue, Oregon State University

We study the scalar-on-function regression model with imprecisely measured values of the predictor function. In this setting, we have a scalar-valued response and a function-valued covariate that are both collected at a single time period. We propose a generalized method of moments-based approach for estimation while an instrumental variable belonging in the same time space as the imprecisely measured covariate is used for model identification. Additionally, no distributional assumption regarding the measurements are assumed, while complex covariance structures are allowed for the measurement errors in the implementation of our proposed methods. In a simulation study, we illustrate that ignoring measurement error leads to biased estimations of the functional coefficient. The simulation studies also confirm our ability to consistently estimate the function-valued coefficient when compared to approaches that ignore potential measurement

errors. Our proposed methods are applied to assess the impact of baseline levels of energy expenditure on BMI among elementary school-aged children.

email: cdtekwe@sph.tamhsc.edu

CALIBRATING VALIDATION SAMPLES WHEN CORRECTING FOR MEASUREMENT ERROR IN INTERVENTION STUDY OUTCOMES

Benjamin Ackerman*, Johns Hopkins Bloomberg School of Public Health
Juned Siddique, Northwestern University Feinberg School of Medicine
Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Many lifestyle intervention trials depend on collecting self-reported outcomes, such as dietary intake, to assess the intervention's effectiveness. Self-reported measures are subject to measurement error, which may impact treatment effect estimation. Methods exist to correct for measurement error using external validation studies, which measure both the self-reported outcome and accompanying biomarker, to model the measurement error structure. However, there is growing concern over the performance of these methods when the validation study differs greatly from the intervention study on pre-treatment covariates that relate to treatment effect. In this paper, we evaluate the impact of such covariate imbalance on measurement error correction methods through simulation, and propose an improvement upon the methods by weighting the validation study using propensity score-type techniques, followed by the implementation of the measurement error correction. We apply the methods to the PREMIER study, a multi-arm lifestyle intervention trial aimed at reducing self-reported sodium intake, and OPEN, a validation study measuring both self-reported diet and urinary biomarkers.

email: backer10@jhu.edu

ASSESSING THE IMPACT OF DIFFERENTIAL MEASUREMENT ERROR ON OUTCOMES IN A LONGITUDINAL LIFESTYLE INTERVENTION STUDY

David A. Aaby*, Northwestern University Feinberg School of Medicine
Juned Siddique, Northwestern University Feinberg School of Medicine

Dietary intervention studies involve repeated assessments over time, typically using self-reported dietary intake. Obtaining accurate measurement of diet and its change over time is a major challenge due to measurement error and can result in biased estimates of the treatment effect. Classical measurement error assumptions may be invalid for longitudinal dietary data from lifestyle intervention trials, as measurement error may be differential. Measurement error may change as a function of time or differ by treatment condition. We use simulations under a variety of scenarios to assess the impact of differential measurement error on the ability to estimate treatment effects. We investigate how different factors influence bias, mean squared error, and coverage of the confidence interval of the treatment effect. We also examine how sample size and the ratio of treatment to controls affects power. Our findings suggest that failing to take into account differential measurement error can influence estimates of the treatment effect.

email: david.aaby@northwestern.edu

COMPARISON OF CAUSAL METHODS FOR AVERAGE TREATMENT EFFECT ESTIMATION ALLOWING COVARIATE MEASUREMENT ERROR

Zhou Feng*, University of Maryland, Baltimore County
Yi Huang, University of Maryland, Baltimore County

In observational studies, propensity score methods are widely used to estimate average treatment effect (ATE). However, it is common in real world data that a covariate is measured with error, which violate the unconfoundedness assumption. Ignoring measurement error and using naive propensity scores estimated by observed covariates will lead to biased ATE estimates. There are only a few causal methods that control the influence of covariate measurement error in ATE estimation, and there is no literature comparing their numerical performances. We conduct systematic simulation studies to compare the methods under rationales with respect to Gaussian vs. binary outcome, continuous vs. discrete underlying true covariate, small vs. large treatment effect, and small vs. large measurement error. The results show that under Gaussian outcome, bias correction method and latent propensity score method using EM algorithm perform best with small and large measurement error respectively; under binary outcome, the inverse probability weighting method and the latent propensity score method using MCMC algorithm perform best with small and large measurement error respectively.

email: zhouf1@umbc.edu

ESTIMATION OF NATURAL INDIRECT EFFECTS ROBUST TO UNMEASURED CONFOUNDING AND MEDIATOR MEASUREMENT ERROR

Isabel R. Fulcher*, Harvard University
Xu Shi, Harvard University
Eric J. Tchetgen Tchetgen, The Wharton School, University of Pennsylvania

The use of causal mediation analysis to evaluate the pathways by which an exposure affects an outcome is widespread in the social and biomedical sciences. Recent advances in this area have established formal conditions for identification and estimation of natural direct and indirect effects. However, these conditions typically involve stringent no unmeasured confounding assumptions and that the mediator has been measured without error. These assumptions may fail to hold in practice where mediation methods are often applied. The goal of this paper is two-fold. First, we show that the natural indirect effect can in fact be identified in the presence of unmeasured exposure-outcome confounding provided there is no additive interaction between the mediator and unmeasured confounder(s). Second, we introduce a new estimator of the natural indirect effect that is robust to both classical measurement error of the mediator and unmeasured confounding of both exposure-outcome and mediator-outcome relations under certain no interaction assumptions. We provide formal proofs and a simulation study to demonstrate our results.

email: isabelfulcher@g.harvard.edu

72. GENOMICS, PROTEOMICS, OR OTHER OMICS

A NOVEL TEST FOR POSITIVE SELECTION USING PROTEIN STRUCTURAL INFORMATION

Peter B. Chi*, Villanova University
David A. Liberles, Temple University

Positive selection refers generically to any process by which an advantageous genetic variant gains increased representation in a population, and its presence is an important consideration when trying to infer relatedness between species. Here, we introduce a new method in which the 3-dimensional structure of proteins is considered, and the statistical question is whether substitutions are occurring physically closer together than expected by chance. Several similar methods have been recently proposed (Adams et al. 2017, Meyer et al. 2016, among others), but ours takes several considerations into account which results in improved performance with respect to Type I Error rate and Power.

email: peter.chi@villanova.edu

PENALIZED MULTIPLE CO-INERTIA ANALYSIS WITH APPLICATION TO INTEGRATIVE ANALYSIS OF MULTI -OMICS DATA

Eun Jeong Min*, University of Pennsylvania
Qi Long, University of Pennsylvania

Multiple co-inertia analysis (mCIA) is a multivariate statistical analysis method for assessing relationships and trends in multiple sets of data. Recently mCIA has been used for integrative analysis of high-dimensional multiple -omics data. However, the estimated loading vectors from the existing mCIA method are non-sparse, presenting challenges for interpreting analysis results. We propose a novel sparse mCIA (smCIA) method that produces sparse estimates and a structured sparse mCIA (ssmCIA) method that enables incorporation of structural information among variables such as those from functional genomics. Extensive simulation studies demonstrate the superior performance of the smCIA and ssmCIA methods compared to the existing mCIA. We also apply our methods to the integrative analysis of transcriptomics data and the proteomics data from a cancer study.

email: mineunj@pennmedicine.upenn.edu

UNDERSTANDING AND PREDICTING RAPID DISEASE PROGRESSION IN THE PRESENCE OF SPARSE EFFECTS: A CASE STUDY WITH CYSTIC FIBROSIS LUNG FUNCTION AND PROTEOMICS DATA

Emrah Gecili*, Cincinnati Children's Hospital and University of Cincinnati
John P. Clancy, Cystic Fibrosis Foundation
Rhonda Szczesniak, Cincinnati Children's Hospital and University of Cincinnati
Assem Ziady, Cincinnati Children's Hospital and University of Cincinnati

Finding reasonable models for understanding and predicting rapid disease progression using high-dimensional data is challenging. For clinical data augmented with large-scale data on proteins, it is not only important to find a model with high predictive accuracy, but also for the model to rely on only a few protein variants and that the selection of these features is stable. Selection methods include the lasso, random forests and PCR; however, each approach has pros and cons regarding sparsity, assessing uniqueness of the variants selected, ease of interpretation and implementation in longitudinal settings. Marginal testing identifies individual variants but lacks efficiency and suffers from low power. A novel hypercube model is used to evaluate explanatory and predictive features of different sets of variables. We implement each selection approach under a Gaussian linear mixed effects model with nonstationary covariance and generate real-time, proteomics-informed predictions of rapid disease progression. We compare these selection approaches in an empirical analysis using proteomics and lung function data observed on individuals with cystic fibrosis lung disease.

email: emrahgecili@gmail.com

MOVIE: MULTI-OMICS VISUALIZATION OF ESTIMATED CONTRIBUTIONS

Sean D. McCabe*, University of North Carolina, Chapel Hill
Dan-Yu Lin, University of North Carolina, Chapel Hill
Michael I. Love, University of North Carolina, Chapel Hill

The growth of multi-omics datasets has given rise to many methods for identifying sources of common variation across data types. The unsupervised nature of these methods makes it difficult to evaluate their performance. We present MOVIE, Multi-Omics VIsualization of Estimated contributions, as a framework for evaluating the degree of overfitting and the stability of unsupervised multi-omics methods. MOVIE plots the contributions of one data type against another to produce contribution plots, where contributions are calculated for each subject and each data type from the results of each multi-omics method. The usefulness of MOVIE is demonstrated by applying existing multi-omics methods to permuted null data and breast cancer data from The Cancer Genome Atlas. Contribution plots indicated that principal components-based Canonical Correlation Analysis overt null data, while Sparse multiple Canonical Correlation Analysis and Multi-Omics Factor Analysis provided stable results with high specificity for both the real and permuted null datasets.

email: mccabes@live.unc.edu

IMPROVED DETECTION OF EPIGENETIC MARKS WITH MIXED EFFECTS HIDDEN MARKOV MODELS

Pedro L. Baldoni*, University of North Carolina, Chapel Hill
Naim U. Rashid, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a technique to detect genomic regions containing protein-DNA interaction, such as transcription factor binding sites or regions containing histone modifications. One goal of the analysis of ChIP-seq experiments is to identify genomic loci enriched for sequencing reads pertaining to DNA bound to the factor of interest. The accurate identification of such regions aids in the understanding of epigenetic marks and gene regulatory mechanisms. Given the reduction in massively parallel sequencing costs, methods to detect consensus regions of enrichment across multiple samples are of interest. Here, we present a statistical model to detect broad consensus regions of enrichment from multiple ChIP-seq experiments through a class of Zero-Inflated Mixed Effects Hidden Markov Models. We show that the proposed model

outperforms existing methods for consensus peak calling in common epigenetic marks by accounting for the excess zeros and sample-specific biases. We applied our method to data from the Encyclopedia of DNA Elements (ENCODE) project and also from an extensive simulation study.

email: baldoni@email.unc.edu

ACCURATE AND EFFICIENT ESTIMATION OF SMALL P-VALUES WITH THE CROSS-ENTROPY METHOD: APPLICATIONS IN GENOMIC DATA ANALYSIS

Yang Shi*, Augusta University
Hui Jiang, University of Michigan
Huining Kang, University of New Mexico Comprehensive Cancer Center
Ji-Hyun Lee, University of Florida

Small p-values are often required to be accurately estimated in large-scale genomic studies for the adjustment of multiple hypothesis tests and the ranking of genomic features based on their statistical significance. For those complicated test statistics whose cumulative distribution functions are analytically intractable, existing methods usually do not work well with small p-values due to lack of accuracy or computational restrictions. We propose a general approach for accurately and efficiently calculating small p-values for a broad range of complicated test statistics based on the principle of the cross-entropy method and Markov chain Monte Carlo sampling techniques. In this talk, I will demonstrate the application of the proposed approach to the accurate and efficient estimation of small p-values in permutation tests and parametric bootstrap tests with real examples in genomic studies.

email: yshi@augusta.edu

SECOND-GENERATION P-VALUES IN A HIGH DIMENSIONAL ANALYSIS OF PROSTATE CANCER VARIANTS

Valerie F. Welty*, Vanderbilt University
Robert A. Greevy, Vanderbilt University
Jeffrey R. Smith, Vanderbilt University
William D. Dupont, Vanderbilt University
Jeffrey D. Blume, Vanderbilt University

The second-generation p-value (SGPV) is a novel extension to the p-value that accounts for scientific relevance by using a composite null hypothesis to capture null and scientifically trivial effects (Blume et al. 2018). The SGPV provides an indication of when the results of a study are compatible with the alternative hypothesis, the null hypothesis, or are inconclusive, and addresses many of the traditional p-value's undesirable properties. A formal definition and introduction will be presented, followed by an illustration of how the second-generation p-value framework can be applied in a high dimensional setting. The SGPV, along with its delta-gap, is used to rank potential associations between 247,000 single-nucleotide polymorphisms (SNPs) on chromosome 6 and the occurrence of prostate cancer. The approach will be contrasted to standard approaches, where the associations are ranked by the (adjusted) classical p-value. The second-generation p-value is seen to have notable advantages over traditional p-values.

email: valerie.welty@vanderbilt.edu

73. IMAGING METHODS

HIERARCHICAL SHRINKAGE PRIORS FOR USING IMAGES AS PREDICTORS IN BAYESIAN GENERALIZED LINEAR MODELS

Justin M. Leach*, University of Alabama at Birmingham
Inmaculada Aban, University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham

Images are increasingly relevant in scientific studies as both predictors and outcomes; e.g. how do MRI scans relate to or predict cognitive decline? Generalized linear models are a classical method for analyzing outcomes whose distributions are exponential families. However, the number of pixels or voxels in an image often exceeds the number of subjects, resulting in a non-identifiable model; further complications may arise from the correlation structure inherent to images, even when the number of subjects exceeds the number of predictors (pixels/voxels). Bayesian approaches have been used in similar contexts, such as genetics, to circumvent these issues by creative specification of prior distributions. Additionally, large image dimensions, often many thousands of pixels or voxels, complicate variable selection and raise concerns about false positives. We explore the application and extension of Bayesian Hierarchical Generalized Linear Models to the problem of using images as predictors, specifically using hierarchical shrinkage priors to identify important predictors, and we explore possible approaches for incorporating the spatial structure of the images into the models.

email: jleach@uab.edu

A SIMPLIFIED CROSSING FIBER MODEL IN DIFFUSION WEIGHTED IMAGING

Kaushik Ghosh*, University of Nevada Las Vegas
Sheng Yang, Case Western Reserve University
Ken Sakaie, Cleveland Clinic
Satya S. Sahoo, Case Western Reserve University
Sarah J. Carr, King's College, London
Curtis Tatsuoka, Case Western Reserve University

Diffusion MRI (dMRI) is a vital source of imaging data for identifying anatomical connections in the living human brain that form the substrate for information transfer between brain regions. Currently, the Ball-and-Stick model serves as a widely implemented probabilistic approach in the tractography toolbox of the popular FSL software package and FreeSurfer/TRACULA software package. However, estimation of the features of neural fibers is complex under the scenario of two crossing neural fibers, which occurs in a sizeable proportion of voxels within the brain. Such models can pose a difficult statistical estimation problem computationally. We propose a simplified version of Ball-and-Stick model that reduces parameter space dimensionality. This simplified model is significantly more efficient in the terms of computation time required in estimating parameters pertaining to two crossing neural fibers through Bayesian simulation approaches. Moreover, the performance of this new model is comparable or better in terms of bias and estimation variance as compared to existing models.

email: kaushik.ghosh@unlv.edu

SPECTRAL ANALYSIS OF BRAIN SIGNALS: A NEW BAYESIAN NONPARAMETRIC APPROACH

Guillermo Cuauhtemoctzin Granados-Garcia*, King Abdullah University of Science and Technology
Mark Fiecas, University of Minnesota
Hernando Ombao, King Abdullah University of Science and Technology

The goal of this paper is to study complex oscillatory behavior of brain signals through their spectral density function (SDF). We develop a Bayesian nonparametric estimation method for SDF. Our approach decomposes the SDF as a mixture of spectral splines where each spectral spline is derived from a unique second-order autoregressive process. Our proposed approach gives easily interpretable results because it decomposes the time series as a discrete mixture of second-order autoregressive processes. Another advantage of our method is that it automatically identifies frequencies of the oscillating components that contribute the most power in the signals. The performance of the proposed model was examined through various realistic simulation settings. The spectral spline approach was used to study human brain recordings to identify significant changes in patterns during sleep with respect to age.

email: guillermo.granadosgarcia@kaust.edu.sa

COPULA MODELING OF SPECTRAL DECOMPOSITIONS OF MULTIVARIATE NON-STATIONARY TIME SERIES

Yongxin Zhu*, King Abdullah University of Science and Technology
Charles Fontaine, King Abdullah University of Science and Technology
Ron Frostig, University of California, Irvine
Hernando Ombao, King Abdullah University of Science and Technology

Brain signals often exhibit non-stationarity as the idiosyncrasies of dependence can change over time. In addition, dependence between brain regions is complex and cannot be fully explained by the classical correlation and coherence-based measures. We develop a new approach to characterize the dependence between time series using copulas. Compared to classical measures, fitting parametric copulas allows us to specify the marginal distributions of each time series, separately from the possible dependence structures among these distributions. This oers greater exibility beyond commonly used distributions in model specication and estimation. In our procedure, brain signals are decomposed into dierent frequency bands and the copulas are used to capture non-linear dual-frequency dependence structures.

email: yongxin.zhu@kaust.edu.sa

COMPARING SUMMARY METHODS AND A SPATIOTEMPORAL MODEL IN THE ANALYSIS OF LONGITUDINAL IMAGING DATA

Brandon J. George*, Thomas Jefferson University
Inmaculada B. Aban, University of Alabama at Birmingham

Summary methods of both spatial and temporal data have been used previously in the analysis of longitudinal imaging data but may be suboptimal due to the potential loss of power associated with the reduction of data. To address this concern, we have previously proposed the use of a linear model with a separable parametric correlation structure for the error terms. A simulation study, whose structure was modeled after longitudinal cardiac imaging data from a clinical trial, was done to compare the statistical properties of our proposed method and several common summary measures in time and space. When testing for a treatment-by-time effect, our model more reliably conserved the Type I error rate and had greater statistical power than the summary methods. The effects of missing data were considered, and the summary measures were not found to improve in relation to our proposed model.

email: bgeorge.bst@gmail.com

A SPATIAL BAYESIAN SEMI-PARAMETRIC MODEL OF POSITIVE DEFINITE MATRICES FOR DIFFUSION TENSOR IMAGING

Zhou Lan*, North Carolina State University
Brian J. Reich, North Carolina State University
Dipankar Bandyopadhyay, Virginia Commonwealth University

Diffusion tensor imaging (DTI) is a popular magnetic resonance imaging technique used to characterize microstructural changes of the brain. Estimated 3x3 positive definite matrices are summarized for each voxel for revealing the structure of the brain. The statistical analysis for DTI data is not convenient due to the difficulty of modeling matrix variate data. In this paper, we propose a spatial Bayesian semi-parametric model of positive definite matrices for modeling the DTI data. The model assumes the positive definite matrix as a mixture of inversed Wishart matrices. The spatial dependency is captured by the Potts model. The nice conjugacy and the implementation of the double Metropolis-Hasting make the computation simple. In simulations, the method has a powerful and robust performance in regions of difference selection. We also apply this method to the cocaine user data. The result is clinically meaningful and shows that the Bayesian semi-parametric model has a good separation property. This property is both desirable in the Bayesian variable selection and clinical studies.

email: zlan@ncsu.edu

74. PRESIDENTIAL INVITED ADDRESS

A PARTICULATE SOLUTION: DATA SCIENCE IN THE FIGHT TO STOP AIR POLLUTION AND CLIMATE CHANGE

Francesca Dominici, Ph.D, Clarence James Gamble Professor of Biostatistics, Population and Data Science, Harvard T.H. Chan School of Public Health and Co-Director, Data Science Initiative, Harvard University

What if I told you I had evidence of a serious threat to American national security – a terrorist attack in which a jumbo jet will be hijacked and crashed every 12 days. Thousands will continue to die unless we act now. This is the question before us today – but the threat doesn't come from terrorists. The threat comes from climate change and air pollution. We have developed an artificial neural network model that uses on-the-ground airmonitoring data and satellite-based measurements to estimate daily pollution levels across the continental U.S., breaking the country up into 1-square-kilometer zones. We have paired that information with health data contained in Medicare claims records from the last 12 years, and for 97% of the population ages 65 or older. We have developed statistical methods and

computational efficient algorithms for the analysis over 460 million health records. Our research shows that short and long term exposure to air pollution is killing thousands of senior citizens each year. This data science platform is telling us that federal limits on the nation's most widespread air pollutants are not stringent enough. This type of data is the sign of a new era for the role of data science in public health, and also for the associated methodological challenges. For example, with enormous amounts of data, the threat of unmeasured confounding bias is amplified, and causality is even harder to assess with observational studies. These and other challenges will be discussed.

email: fdominic@hsph.harvard.edu

75. RESOURCE EFFICIENT STUDY DESIGNS FOR OBSERVATIONAL AND CORRELATED DATA

MULTI-WAVE, RESPONSE-SELECTIVE STUDY DESIGNS FOR LONGITUDINAL BINARY DATA

Nathaniel D. Mercaldo*, Massachusetts General Hospital and Harvard University
Jonathan S. Schildcrout, Vanderbilt University Medical Center

Retrospective outcome dependent sampling (ODS) designs are an efficient class of study designs that may be implemented when resource constraints prohibit ascertainment of an expensive covariate on all members of a cohort. One class of ODS designs for longitudinal binary data stratifies individuals into strata according to those who never, sometimes, and always experience the binary event. If inference lies in a time-invariant covariate effect, or in the joint effect of time-varying and time-invariant covariates, then the design choice is not clear. Since the ideal design for many estimation targets is not always obvious, we propose a class of multi-wave ODS designs for longitudinal binary data where later wave designs are identified after data have been collected and examined at earlier waves. We will describe a class of two-wave designs, examine their finite sampling operating characteristics, and apply the designs to an exemplar longitudinal cohort study, the Lung Health Study.

email: nmercaldo@mgh-ita.org

EFFICIENT DESIGN AND ANALYSIS OF TWO-PHASE STUDIES WITH A LONGITUDINAL CONTINUOUS OUTCOME

Ran Tao*, Vanderbilt University Medical Center

In modern longitudinal studies, the covariates of interest may involve genetic profiling, biomarker assay, or medical imaging and thus are prohibitively expensive to measure on a large number of subjects. A cost-effective solution is the two-phase design, under which the longitudinal outcome trajectory and inexpensive covariates are observed for all subjects during the first phase and this information is used to select subjects for measurements of expensive covariates during the second phase. Herein, we consider general two-phase designs with longitudinal continuous response data, where the sampling in the second phase can depend on the first-phase data in any manner, and the inexpensive covariates can be continuous and correlated with expensive covariates. We propose a semiparametric approach to regression analysis and establish the consistency, asymptotic normality, and asymptotic efficiency of the estimators. In addition, we derive optimal two-phase designs, which can be substantially more efficient than existing designs. Finally, we demonstrate the usefulness of the proposed designs and inference procedures through simulation studies and real-world applications.

email: r.tao@vanderbilt.edu

RESPONSE DRIVEN STUDY DESIGNS FOR MULTIVARIATE LONGITUDINAL DATA

Jonathan S. Schildcrout*, Vanderbilt University Medical Center

In an era of 'big data' warehouses, efficient study designs and analytical procedures are vital. Towards this, response-driven study designs (RDSD) have had an enormous impact because they permit the use of these data resources to address hypotheses that could not otherwise be investigated. The most common RDSD is the case-control design although other outcome dependent sampling (ODS) and outcome-related, auxiliary variable dependent sampling (AVDS) schemes have been also been developed. In prior research we showed that outcome dependent sampling designs can be highly efficient for parameters describing exposure-outcome relationships in longitudinal data settings. In this talk, we consider the scenario when analysis of more than one longitudinal outcome is of interest. Specifically, we extend our previous work to describe a class of designs that sample based on multiple longitudinal response vectors. We will describe the class of designs and likelihood-based analysis procedures. As an exemplar we examine genetic modification of lung function decline in a cohort of patients with chronic obstructive pulmonary disease.

email: jonathan.schildcrout@vanderbilt.edu

SEMIPARAMETRIC GENERALIZED LINEAR MODELS: APPLICATION TO BIASED SAMPLES

Paul Rathouz*, Dell Medical School, University of Texas, Austin

Rathouz and Gao (2009) proposed a novel class of generalized linear models indexed by a linear predictor and a link function for the mean of (Y|X). In this class, the distribution of (Y|X) is left unspecified and estimated from the data via exponential tilting of a reference distribution, yielding a response model that is a member of the natural exponential family. We provide a brief overview of theoretical and computational developments, covering both finite and infinite support and small and large samples. We then focus on how, with very easy-to-implement modifications, the model can accomodate biased samples arising from extensions of case-control designs to continuous response distributions, wherein inferences about the mean are of interest.

email: paul.rathouz@austin.utexas.edu

76. RECENT ADVANCES IN THE STUDY OF INTERACTION

THE INTERACTION CONTINUUM

Tyler J. VanderWeele*, Harvard University

A common reason for assessing interaction is to evaluate "whether the effect is larger in one group vs. another". However, the answer to this question is scale dependent.

It is shown the joint outcomes probabilities can be placed on "interaction continuum." When both main effects for two exposures are positive then the placement on the continuum depends upon the relative magnitude of the probability of the outcome in the doubly exposed group. For high probabilities of the outcome in the doubly exposed group the interaction may be positive-multiplicative positive-additive. As the probability of the outcome in the doubly exposed group goes down, the form of interaction descends through the ranks of: positive-multiplicative positive-additive; no-multiplicative positive-additive, negative-multiplicative positive-additive, negative-multiplicative zero-additive, negative-multiplicative negative-additive, single pure interaction, single qualitative interaction, single-qualitative single-pure interaction, double qualitative interaction, perfect antagonism, inverted interaction.

email: tvanderw@hsph.harvard.edu

ADDITIVE VERSUS MULTIPLICATIVE INTERACTION: THE EPIDEMIOLOGICAL FOLKLORE REGARDING HETEROGENEITY ACROSS STUDIES

Bhramar Mukherjee*, University of Michigan

Heuristically and empirically it has been observed and claimed that the risk difference estimates are more heterogeneous across studies than the risk ratio estimates in a meta-analysis setting. We study this phenomenon for interactions instead of marginal effects under both cohort and case-control design and arrive at the conclusion that study design and how you define "heterogeneity" has an effect on this claim. By studying the asymptotic distribution of Cochran's Q statistic for testing homeogeneity, we justify our claim when the number of studies is K=2, and simulation results support the plausibility of our results for K>2. This is joint work with Sehee Kim and Greyson Liu.

email: bhramar@umich.edu

A GENERAL APPROACH TO DETECT GENE (G)-ENVIRONMENT (E) ADDITIVE INTERACTION LEVERAGING G-E INDEPENDENCE IN CASE-CONTROL STUDIES

Xu Shi*, Harvard University
Eric J. Tchetgen Tchetgen, Wharton School of the University of Pennsylvania
Tamar Sofer, Brigham and Women's Hospital and Harvard Medical School
Benedict H. W. Wong, Harvard University

In case-control studies involving a rare disease, a statistical test of no additive interaction between genetic (G) and environmental (E) typically entails a test of no relative excess risk due to interaction (RERI). It has been shown that a likelihood ratio test of a null RERI incorporating the G-E independence assumption (RERI-LRT) outperforms the standard RERI approach. The RERI-LRT relies on correct specification of a logistic model for the binary outcome, as a function of G, E and auxiliary covariates. However, when at least one exposure is non-discrete or covariates are present, nonparametric estimation may not be feasible, and parametric logistic regression will a priori rule out the null hypothesis of no additive interaction, inflating the type 1 error rate. In this talk, we present a general approach to test for additive interaction exploiting G-E independence. It avoids specification of an outcome model, but still allows for covariate adjustment in separate regression models for G and E to ensure the G-E independence assumption or to rule out residual confounding. The methods are illustrated through an extensive simulation study and an ovarian cancer application.

email: xushi@hsph.harvard.edu

77. EXPANDING RANK TESTS: ESTIMATES, CONFIDENCE INTERVALS, MODELING, AND APPLICATIONS

CONFIDENCE INTERVALS AND CAUSAL INFERENCE WITH THE MANN-WHITNEY PARAMETER THAT ARE COMPATIBLE WITH THE WILCOXON-MANN-WHITNEY TEST

Michael P. Fay*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Yaakov Malinovsky, University of Maryland, Baltimore County
Erica H. Brittain, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Joanna H. Shih, National Cancer Institute, National Institutes of Health
Dean A. Follmann, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Erin E. Gabriel, Karolinska Institute, Stockholm, Sweden

The Wilcoxon-Mann-Whitney (WMW) test is used frequently for the two-sample problem with numeric or ordinal responses. Unfortunately, the WMW test is rarely presented with an effect estimate and confidence interval. The natural effect estimate is the Mann-Whitney parameter, Pr[Y1<Y2]+0.5Pr[Y1=Y2]. We present two recent innovations with the MW parameter related to the WMW test. First, we show that the MW parameter may be represented as a causal effect and is estimable with minimal assumptions from a randomized experiment. Second, we develop a different confidence interval for the MW parameter that is compatible with each of the common implementations of the WMW test, including the exact implementation and the asymptotically normal implementation with a continuity correction. Compatible in this case means that a WMW test rejects at level alpha if and only if the 1-alpha confidence interval for the MW parameter associated with that implementation excludes 1/2. We discuss assumptions, interpretation, and practical implementation of the tests and confidence intervals using our asht R package.

email: mfay@niaid.nih.gov

SMALL SAMPLE INFERENCE FOR PROBABILISTIC INDEX MODELS

Jan De Neve*, Ghent University
Gustavo Amorim, Ghent University
Olivier Thas, Ghent University, Hasselt University and University of Wollongong
Karel Vermeulen, Ghent University
Stijn Vansteelandt, Ghent University and London School of Hygiene and Tropical Medicine

We demonstrate how many classical rank tests, such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis and Friedman test, can be embedded in a statistical modelling framework and how the method can be used to construct new rank tests. In addition to hypothesis testing, the method allows for estimating effect sizes with an informative interpretation, resulting in a better understanding of the data. Our method results from two particular parametrizations of probabilistic index models (Thas et al., 2012). The popularity of rank tests for small sample inference makes probabilistic index models also natural candidates for small sample studies. However inference for such models relies on asymptotic theory that can deliver poor approximations of the sampling distribution if the sample size is rather small. We therefore explore a bias-reduced version of the bootstrap and adjusted jackknife empirical likelihood and show that their application leads to drastic improvements in small sample inference

for probabilistic index models. These results justify the use of such models for reliable and informative statistical inference in small sample studies.

email: JanR.DeNeve@ugent.be

RANK-BASED PROCEDURES IN FACTORIAL DESIGNS: HYPOTHESES ABOUT NONPARAMETRIC TREATMENT EFFECTS

Frank Konietschke*, University of Texas, Dallas

Existing tests for factorial designs in the nonparametric case are based on hypotheses formulated in terms of distributions. Typical null hypotheses, however, are formulated in terms of some parameters or effect measures, particularly in heteroscedastic settings. In this talk we extend this idea to nonparametric models by introducing a novel nonparametric ANOVA-type-statistic based on ranks which is suitable for testing hypotheses formulated in meaningful nonparametric treatment effects in general factorial designs. This is achieved by a careful in-depth study of the common distribution of rank-based estimators for the treatment effects. Since the statistic is asymptotically not a pivotal quantity we propose different approximation techniques, discuss their theoretic properties and compare them in extensive simulations together with two additional Wald-type tests. An extension of the presented idea to general repeated measures designs is briefly outlined. The proposed rank-based procedures maintain the pre-assigned type-I error rate quite accurately, also in unbalanced and heteroscedastic models. A real data example illustrates the application of the proposed methods.

email: fxk141230@utdallas.edu

DESIRABILITY OF OUTCOME RANKING (DOOR): MOTIVATION AND EXAMPLES

Scott R. Evans*, The George Washington University

RCTs are the gold standard for evaluating the benefits and risks of interventions. However trials often lack pragmatism, failing to provide the necessary evidence to inform practical medical decision-making. The implications of these deficiencies are largely absent from discourse in medical research. Typically a treatment effect on each outcome of interest is estimated and the effects are formally or informally combined. However assessment based on combining the separate marginal effects on each outcome carries several limitations. The desirability of outcome ranking (DOOR) is an evolving approach based on pairwise comparisons of patients between randomized treatment arms, providing a quantitative global assessment that addresses these challenges. The concept uses the outcomes to analyze patients rather than patients to analyze outcomes, accommodating several outcomes with differential importance. Crucial to this approach entails improved understanding of how to analyze one patient before analyzing many. Ranking global patient outcomes provides a more clinically relevant evaluation to inform clinical decision-making than combining separate marginal effects.

email: sevans@bsc.gwu.edu

78. NOVEL STATISTICAL METHODS TO ANALYZE SELF-REPORTED OUTCOMES SUBJECT TO RECALL ERROR IN OBSERVATIONAL STUDIES

ANALYZING RECURRENT EVENTS IN PRESENCE OF RECALL ERROR: AN APPLICATION TO TIME-TO-HOSPITALIZATION

Rajeshwari Sundaram*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Sedigheh Mirzaei, St. Jude Children's Research Hospital
Edwina Yeung, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Measurement error in time to event data used as a predictor leads to inaccurate inference. Typically, time-to-event predictor measured with error are often encountered in the context of self-reported family history. This issue also occurs in recall of self-reported recurrent events like repeated hospitalizations. Additionally, such recall information is typically collected intermittently through repeated surveys or patient recalls over time. Thus, resulting in panel count data with either error in number of events recalled or the duration of repeated events recalled erroneously or both. Using a validation data set, we propose a method to adjust for this type of measurement error. We propose a joint modeling framework that accounts for the dependence between the measurement error and the underlying recurrent event data. We use the estimated measurement error model to adjust the likelihood for the underlying recurrent events data. The proposed method is investigated through simulations and are illustrated by analyzing the time to (repeated) hospitalization of infant/child data of Upstate KIDS study.

email: sundaramr2@mail.nih.gov

ESTIMATION OF MENARCHEAL AGE DISTRIBUTION FROM IMPERFECTLY RECALLED DATA

Debasis Sengupta*, Indian Statistical Institute, Kolkata
Sedigheh Mirzaei Salehabadi, St. Jude Children's Research Hospital
Rahul Ghosal, North Carolina State University

In a cross-sectional retrospective study on menarcheal age, some respondents who had experienced menarche were able to recall the date exactly, some recalled only the month or year of the event, and some were unable to recall it. This interval censored data bears evidence of being informatively censored, for which we build a special model for estimating the menarcheal age distribution. We provide a set of regularity conditions on the distribution, subject to which the consistency and the asymptotic normality of the parametric maximum likelihood estimator are established. We also provide a computationally simple approximation to the non-parametric maximum likelihood estimator and establish its consistency under mild conditions. We study the small sample performance of the two estimators through Monte Carlo simulations. Moreover, we provide a graphical check for the assumption of the multinomial regression model for the recall probabilities. The assumption appears to hold for the menarcheal data set. Its analysis shows that the use of the

imperfectly recalled part of the data in the proposed manner indeed leads to smaller confidence intervals of the survival function.

email: sdebasis@isical.ac.in

VARIABLE SELECTION IN HIGH DIMENSIONAL DATASETS IN THE PRESENCE OF SELF-REPORTED OUTCOMES

Raji Balasubramanian*, University of Massachusetts Amherst
Mahlet G. Tadesse, Georgetown University
Andrea S. Foulkes, Mount Holyoke College
Yunsheng Ma, University of Massachusetts Medical School
Xiangdong Gu, University of Massachusetts Amherst

The onset of silent diseases such as type 2 diabetes are often ascertained through self-report in large prospective cohorts. While cost effective, self-reported outcomes are subject to error. In other settings, diagnosis of silent events may occur through the use of an imperfect diagnostic test. Here, we adapt a Bayesian variable selection approach to address the goal of variable selection in high dimensional datasets, when the time to event outcome is ascertained with error. We perform simulation studies to assess the performance of our proposed method and compare to a naive approach that ignores error in the outcome ascertainment. We apply our proposed method to discover single nucleotide polymorphisms (SNPs) that are associated with risk of type 2 diabetes in a dataset of 12,008 women in the Women's Health Initiative.

email: rbalasub@umass.edu

MEASUREMENT ERROR CORRECTION IN LONGITUDINAL DIETARY INTERVENTION STUDIES IN THE PRESENCE OF NONIGNORABLE MISSING DATA

Juned Siddique*, Northwestern University
Caroline P. Groth, Northwestern University
David A. Aaby, Northwestern University
Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health
Michael J. Daniels, University of Florida

In lifestyle intervention trials, where the goal is to change a participant's weight or modify their eating behavior, self-reported diet is a longitudinal outcome variable that is subject to measurement error. An additional challenge is participant drop out due to the intensive nature of the interventions that include self-monitoring, following a prescribed diet, coaching, and/or increasing physical activity. We propose a statistical framework for correcting for measurement error in longitudinal self-reported dietary data by combining intervention data with auxiliary data from an external biomarker validation study where both self-reported and recovery biomarkers of dietary intake are available. In this setting, dietary intake measured without error in the intervention trial is missing data and multiple imputation is used to fill in the missing measurements. We use sensitivity analyses to address the influence of unverifiable assumptions involving dropout and its association with the measurement error process. We apply our methods to self-reported sodium intake from the PREMIER study, a multi-component lifestyle intervention trial.

email: siddique@northwestern.edu

79. STATISTICAL ADVANCE IN HUMAN MICROBIOME DATA ANALYSIS

TESTING HYPOTHESES ABOUT ASSOCIATIONS BETWEEN TAXONOMIC GROUPINGS AND TRAITS USING 16S rRNA DATA: A BOTTOM-UP APPROACH

Glen A. Satten*, Centers for Disease Control and Prevention
Yijuan Hu, Emory University
Yunxiao Li, Emory University

Microbiome association tests that consider the effect of individual operational taxonomic units (OTUs) typically report marginal associations for each feature. Here we consider the question of whether groups of OTUs that correspond to taxonomic ranks (species, genus, family etc.) are associated with the trait under study. These tests have the dependence structure of the phylogenetic tree describing the bacteria observed in the study. Current tree-structured tests use a top-down approach. For the microbiome, this would correspond to first testing kingdom, then phylum etc., stopping when a non-significant result is obtained. Our view is that a bottom-up approach that tests species-level association, then genus level, etc. makes more scientific sense. Thus, we give a bottom-up testing algorithm that controls the error rate of decisions made at higher levels in the tree, conditional on findings at lower levels in the tree. By simulation, we also show that our approach is better at finding the highest level taxon having the property that all taxonomic groups below the detected taxon are associated with the trait of interest.

email: gas0@cdc.gov

BETA-DIVERSITY DISCRIMINATORY POWER: COMPARISON OF PERMANOVA, MIRKAT, AND USING STANDARD MICROBIOME REFERENCE GROUPS

Mitchell Henry Gail*, National Cancer Institute, National Institutes of Health
Yunhu Wan, National Cancer Institute, National Institutes of Health

Comparisons of microbiome communities are often based on pairwise beta-diversity measures, such as Bray-Curtis. For two-group discrimination, a permutation test (PERMANOVA) and kernel expansion test (Mirkat) are used. A recent proposal computes mean distance from a microbiome test sample to a set of standard reference samples. If several different types of reference samples are used, each test sample has a vector of mean distances. Hotelling's T-squared test can be used on these vectors. We compare the discriminatory power of PERMANOVA, Mirkat and use of standard reference groups. We used 16S data from the Human Microbiome Project for the groups to be discriminated (e.g. skin versus saliva samples). The two groups are easy to discriminate, but by using mixtures, we create a sequence of increasingly difficult discrimination problems. Simulations determine power. Although each of these methods shows good discriminatory power, preliminary data show that the reference group method performs as well or better than the more complex methods for Bray-Curtis distance. Additional studies are needed for other beta-diversity measures.

email: gailm@mail.nih.gov

ZERO-INFLATED LOGISTIC NORMAL MODEL FOR ANALYZING MICROBIOME RELATIVE ABUNDANCE DATA

Zhigang Li*, University of Florida

Massive high dimensional human microbiome data is commonly seen in molecular epidemiology research and have substantially increased in complexity to address critical health concerns due to complex data structure. Analysis challenges arise from compositional, phylogenetically hierarchical, sparse and high dimensional structure of microbiome data. Compositional structure could induce spurious relationships due to the linear dependence between compositional components. In addition, the hierarchical structure of microbiome data from the phylogenetic tree generates dependence at the hierarchical levels which poses a further modeling challenge. Furthermore, the sparsity of microbiome data due to excessive zero sequencing reads for microbial taxa remains an unresolved issue in the literature. Coupled with the high dimensional feature, microbiome data raises great challenging problems in the field of mediation data analysis. We will develop a zero-inflated logistic normal model to address these issues. A simulation study will show the performance of the approach and a real study example will be included as well.

email: zhigang.li@ufl.edu

INTERACTIVE STATISTICAL ANALYSIS OF HOST-MICROBIOME INTERACTION

Hector Corrada Bravo*, University of Maryland, College Park

Community-wide profiling of microbiomes in human host environments concomitant to a variety of measurements of host response via transcriptomics or epigenomics allows researchers to piece together the intricate interaction between host and the microbiome. Integration and analysis of these measurements requires sophisticated statistical analysis with a multitude of assumptions, decisions and inferences. We argue that tight integration of these methods with interactive visualization allows data analysts to better navigate the complexities of these statistical modeling, examine the assumptions and consequences of modeling decisions and ultimately increase or decrease their and their collaborator's confidence in results. In this talk we will introduce both statistical methodology for integrative analysis of microbiome and host data, along with software tools for interactive visualization that are tightly integrated with these statistical methods.

email: hcorrada@umiacs.umd.edu

80. STATISTICAL MEDIATION ANALYSIS FOR HIGH-DIMENSIONAL DATA

LEARNING CAUSAL NETWORKS VIA ADDITIVE FAITHFULNESS

Kuang-Yao Lee*, Temple University
Tianqi Liu, Google LLC
Bing Li, The Pennsylvania State University
Hongyu Zhao, Yale University

In this work we introduce a statistical model, called additively faithful directed acyclic graph (AFDAG), for causal learning from observational data. Our approach is based on additive conditional independence (ACI), a recently proposed three-way statistical relation that shares many similarities with conditional independence but without resorting to multivariate kernels. This distinct feature strikes a balance between a parametric model and a fully nonparametric model, which makes the proposed model attractive for handling large networks. We develop an estimator for AFDAG based on a linear operator that characterizes ACI, and establish the consistency and convergence rates of this estimator, as well as the uniform consistency of the estimated DAG. Moreover, we introduce a modified PC-algorithm to implement the estimating procedure efficiently, so that its complexity is determined by the level of sparseness rather than the dimension of the network. The usefulness of AFDAG formulation is demonstrated through both synthetic examples, and an application to a proteomics data set.

email: kuang-yao.lee@temple.edu

STATISTICAL METHODOLOGY FOR HIGH DIMENSIONAL MEDIATION ANALYSIS

Andriy Derkach*, National Cancer Institute, National Institutes of Health
Joshua Sampson, National Cancer Institute, National Institutes of Health

We describe statistical methods for mediation analysis in epidemiologic studies with an exposure, a high dimensional set of biomarkers and an outcome. We consider two scenarios. In the first scenario, we assume the biomarkers are apriori grouped into pathways and, in the second scenario, we assume there is aprior information about the biomarkers. In the first scenario, we offer a two-step approach for identifying mediating pathways. In the first step, we select groups of biomarkers showing nominal associations with both the exposure and outcome, and in the second step, we identify specific biomarkers within those candidates. In the second scenario, we assume that the exposure directly influences a group of latent, or unmeasured, factors which are associated with both the outcome and a subset of biomarkers. The biomarkers associated with the latent factors linking the exposure to the outcome are "mediators". We derive and develop maximization algorithm to L1-penalized version of the likelihood for this model to limit the number of factors and associated biomarkers. We demonstrate that these new procedures can have higher power for detecting mediators in simulation and data example.

email: andriy.derkach@nih.gov

HYPOTHESIS TESTING IN HIGH-DIMENSIONAL INSTRUMENTAL VARIABLES REGRESSION WITH AN APPLICATION TO GENOMICS DATA

Jiarui Lu*, University of Pennsylvania
Hongzhe Li, University of Pennsylvania

Studies have shown that gene expressions and genetic variants are associated with common human diseases. However, the association can be impacted by potential unmeasured confounders from multiple sources. Since genetic variants explain gene expression variations, they can be used as instruments in the framework of high dimensional instrumental variable regression. However, because the dimensions of both genetic variants and gene expressions are often large, statistical inferences for such high dimensional IV models are not trivial and have not been investigated in literature. This paper considers the hypothesis testing for sparse IV regression models and presents methods for testing regression coefficient, where the test statistic for each single coefficient is constructed based on an inverse regression. A multiple testing procedure is developed for selecting variables and is shown to control the false

discovery rate. Simulations are conducted to evaluate the performance of our methods. These methods are illustrated by an analysis of yeast dataset in order to identify genes that are associated with growth in the presence of hydrogen peroxide.

email: jiaruilu@pennmedicine.upenn.edu

MEDIATION ANALYSIS IN GENOMIC SETTINGS IN THE PRESENCE OF REVERSE CAUSALITY AND OTHER CHALLENGES

Joshua Millstein*, University of Southern California

Multiscale omics studies have the potential to identify molecular components of disease pathways. Mediation analysis is an increasingly important part of this effort by distinguishing between competing causal models. However, in genomic settings challenges such as pleiotropy, reverse causality, unmeasured confounding, and measurement error violate assumptions of many methods. The causal inference test (CIT) has been found useful in genomic settings, but there is some concern about its performance in the presence of unmeasured confounding or measurement error. In simulation studies, I show that that if pleiotropy occurs with unmeasured confounding, type I error is inflated for the CIT and other methods. A new non-centrality parameter mitigates this problem without substantially affecting power. In a high-dimensional testing setting, this version of the CIT is more robust than a variety of other mediation analysis approaches to reverse causality, pleiotropy with and without unmeasured confounding, and reverse causation with measurement error. Also, in the presence of measurement error, it performs favorably in terms of power.

email: joshua.millstein@usc.edu

81. PREDICTION AND PROGNOSTIC MODELING

DYNAMIC RISK PREDICTION USING LONGITUDINAL BIOMARKERS OF HIGH DIMENSIONS

Lili Zhao*, University of Michigan
Susan Murray, University of Michigan

Longitudinal data from biomarkers are often collected in studies and these repeatedly measured variables provide important information regarding the probability of an outcome of interest occurring at a future time. As new technologies have become available for the biomarker discovery, a large number of biomarker measurements are now available over time to study disease progression. A large amount of data provides a more complete picture of the disease progression of a patient and allows us to make more accurate predictions, but it also dramatically increases the difficulties in the statistical modelling. Existing approaches suffer immensely from the curse of dimensionality. In this study, we develop methods for making dynamic risk predictions using repeatedly measured biomarkers of a high dimension, including cases when the number of markers is larger than the sample size. The proposed methods are computationally simple yet sufficiently flexible to capture complex relationships between longitudinal biomarkers and events times. The proposed approaches were evaluated by simulation studies and applied to a dataset from the Nephrotic Syndrome Study Network.

email: zhaolili@umich.edu

ESTIMATING DISEASE ONSET FROM CHANGE POINTS OF MARKERS MEASURED WITH ERROR

Unkyung Lee*, Texas A&M University
Raymond J. Carroll, Texas A&M University
Karen Marder, Columbia University Medical Center
Yuanjia Wang, Columbia University
Tanya P. Garcia, Texas A&M University

Huntington disease is an autosomal dominant, neurodegenerative disease, without clearly identified biomarkers for clinical diagnosis. Current standards for clinical diagnosis rely on a clinician's subjective judgment that a patient's extrapyramidal signs are unequivocally associated with HD. The subjectivity is conducive to error, so that a data–driven, objective definition is needed. Recent studies of motor–sign decline have revealed that disease–onset is closely related to an inflection point of the symptom progression trajectory. We propose a nonlinear location–shift marker model that captures disease symptom progression and assesses how the progression depends on different potential biomarkers. We develop a multi–stage nonparametric estimation procedure to estimate the disease symptom progression and its inflection point. In an empirical study, our estimator is shown to be more robust to model misspecification. Applying our estimator to PREDICT–HD, a large observational study of HD, leads to earlier prediction of receiving an objective HD motor–diagnosis as compared to the current subjective clinical diagnosis.

email: unkyunglee@stat.tamu.edu

DEVELOPMENT OF AN INTERNATIONAL PROSTATE CANCER RISK TOOL INTEGRATING DATA FROM MULTIPLE HETEROGENEOUS COHORTS

Donna Pauler Ankerst*, Technical University Munich
Johanna Straubinger, Technical University Munich

The Prostate Biopsy Collaborative Group (PBCG) collects prostate cancer risk factors and biopsy outcomes from multiple international centers with the objective to improve biopsy decision-making. We recently developed a global risk prediction model for outcomes on prostate biopsy, and posted it online for worldwide use. Construction of the logistic regression model required integrating data from 8492 biopsies from ten diverse centers across North America and Europe. It involved decisions on choice of model, for example, whether to include random effects, as well as the detection of outlying cohorts. We outline here the data visualization and cohort-permutation internal validation strategy used to develop the PBCG model, the utility of which makes it applicable for future multi-cohort modeling applications.

email: ankerst@tum.de

PREDICTING SERVICE USE AND FUNCTIONING FOR PEOPLE WITH FIRST EPISODE PSYCHOSIS IN COORDINATED SPECIALTY CARE

Melanie M. Wall*, Columbia University
Jenn Scodes, Columbia University
Cale Basaraba, Columbia University

A key initiative in research focused on treatment for first episode psychosis (FEP) is improving the implementation of evidence-based coordinated specialty care

(CSC). One area of improvement is expected to come from improved data analytics facilitated by linking different clinical sites through common data elements and a unified informatics approach for aggregating and analyzing patient level data. The present study examines to what extent predictive modeling of patient-level outcomes based on background variables collected at intake and throughout care can be used to differentiate individuals in a way that is useful. Using data from 600 FEP patients from 15 different CSC sites, we will develop and compare several machine learning models for predicting multivariate, correlated outcomes across one year of care. Presentation of results will focus on interpretability of differential prediction across sites and usefulness for facilitating service decisions.

email: mmwall@columbia.edu

RACE AND GENDER SPECIFIC STATISTICAL COMPARISON OF GROWTH CURVE MODELS IN FIRST YEARS OF LIFE

Mehmet Kocak* The University of Tennessee Health Science Center
Alemayehu Wolde, University of Memphis
Frances A. Tylavsky, The University of Tennessee Health Science Center

In pediatric research, anthropometric measures at a specific age may be required for certain developmental assessments such as energy expenditure. This necessitates the choice of a growth model with desired prediction accuracy. To provide guidance with this issue, we compared Logistic and Gompertz growth models, using different sets of parameterization in a race and gender-specific fashion on growth data from a prospective birth-cohort study (CANDLE). We compared the competitive models in terms of absolute residuals and prediction errors, for height, weight, and head circumference. We have shown that the Gompertz model with only the first parameter defined with a subject-specific random effect is the best model in terms of prediction accuracy. Although the same Gompertz model fitted on each individual profile without a random effect also has similar prediction accuracy, it has inflated standard error of estimation as expected, thus, not recommended to be used. We conclude that Gompertz model with only the first parameter defined with a random effect performs the best for height, weight, and head circumference growth in the first four years of life.

email: mkocak1@uthsc.edu

82. ADAPTIVE DESIGNS FOR CLINICAL TRIALS

VALIDITY AND ROBUSTNESS OF TESTS IN SURVIVAL ANALYSIS UNDER COVARIATE-ADAPTIVE RANDOMIZATION

Ting Ye*, University of Wisconsin, Madison
Jun Shao, University of Wisconsin, Madison

Covariate-adaptive (CA) randomization is popular in clinical trials with sequentially arrived patients for balancing treatment assignments across prognostic factors. However, there exists no theoretical work about testing hypotheses under CA randomization in survival analysis, although CA randomization has been used for a long time and its main application is in survival analysis. Often times, practitioners would simply adopt a conventional test such as the log-rank test or score test to compare two treatments, which is controversial. In this article, we prove that the log-rank test is conservative in terms of type I error under CA randomization, and the robust score test developed under simple randomization is no longer robust under CA randomization. We then propose a calibration type log-rank or score test that is valid and robust under CA randomization. Furthermore, we obtain Pitman efficacy of log-rank and score tests to compare their asymptotic relative efficiency. Simulation studies about the type I error and power of various tests are presented under several popular randomization schemes.

email: tye27@wisc.edu

PLATFORM TRIAL DESIGNS FOR BORROWING ADAPTIVELY FROM HISTORICAL CONTROL DATA

James Paul Normington*, University of Minnesota

We propose a Bayesian adaptive platform trial design that uses commensurate prior methods at interim analyses to borrow adaptively from the control group of an earlier-starting trial to reduce the size of the current trial control arm. The design adjusts the trial's randomization ratio in favor of the novel treatment when the interim posterior indicates commensurability of the two control groups. In this setting, our design supplements a control arm with historical data and randomize more new patients to the novel treatments. This design is both ethical and economical, since it shortens the process of introducing new treatments into the market and any additional costs introduced by this design will be compensated by the savings in control arm sizes. Our approach performs well across settings with varying degrees of commensurability and treatment effects, and compares favorably to an all-or-nothing approach in which the decision to pool or discard historical controls is based on a simple ad-hoc frequentist test at interim analysis. We finally consider a three drug extension where a new intervention joins the platform, and show that this also performs well via simulation.

email: jpnormington@gmail.com

IMPROVED ESTIMATES FOR RECIST RESPONDER RATES IN THE SMALL TREATMENT ARMS IN PLATFORM SCREENING TRIALS

James Dunyak*, Astrazeneca
Nidal Al-Huniti, Astrazeneca

Platform trials improve screening efficiency and accelerate drug development by rapidly testing multiple treatments by estimating overall (partial or complete) response rates. Although the underlying tumor load measurements are continuous, clinical project teams usually count responders using RECIST criteria, counting subjects with best response of at least 30% shrinkage from baseline. This discretization loses information, substantially increasing confidence interval (CI) widths and reducing decision quality. We develop a method to directly estimate P(best shrinkage > 30%) through a continuous model of tumor shrinkage [Wild et al, ENVIRONMETRICS, VOL. 7, 1996]. We test the new method using randomly drawn cohorts (size 7) from a large Phase 3 study, with a responder rate of 0.78 and use 80% confidence intervals (CI). The new method has superior and better balanced CI coverage: P(0.78 < CI)=0.10 and P(0.78>CI)=0.08. Counting responders, the Exact binomial CI has poor coverage: P(0.78 < CI)<0.01 and P(0.78>CI)=0.04. Counting-based, the Agresti-Coull binomial has better coverage but is poorly balanced: P(0.78 < CI)<0.01 and P(0.78>CI)=0.16.

email: james.dunyak@astrazeneca.com

A CONTINUOUS REASSESSMENT METHOD FOR PEDIATRIC PHASE I CLINICAL TRIALS

Yimei Li*, University of Pennsylvania
Ying Yuan, University of Texas MD Anderson Cancer Center

Pediatric phase I trials are usually carried out after the adult trial testing the same agent already started and accumulated some toxicity data. As the pediatric trial progresses, when new data from the concurrent adult trial becomes available, amendments of the pediatric protocol may be submitted to modify the original dose escalation design based on the updated adult trial information. We aim to develop a dose finding design that could systematically incorporate concurrent adult trial data into the pediatric dose finding procedure. We propose a CRM that uses a discounted joint likelihood of the adult and pediatric data with a discount parameter to reflect the belief about the degrees of congruence between pediatric and adult data. The discount parameter is assigned a discrete probability mass which is then updated as the trials progress. The posterior probabilities of toxicity are estimated by the Bayesian model averaging approach and these estimates are used in dose finding procedure for the pediatric trial. Through simulation studies, we demonstrate that the proposed method has higher probability to select the true MTD and is robust to various assumptions.

email: liy3@email.chop.edu

A SIMULATION-BASED SAMPLE SIZE DETERMINATION FOR ADAPTIVE SEAMLESS PHASE II/III DESIGN

Zhongying Xu*, University of Pittsburgh
John A. Kellum, University of Pittsburgh
Gary M. Marsh, University of Pittsburgh
Chung-Chou H. Chang, University of Pittsburgh

The adaptive seamless phase II/III design combines conventional separate phases II and III trials into a single one, and it allows adaptations (e.g. sample size reassessment and early stopping for futility or success) after the interim analysis. In this study, we propose a simulation-based method to determine the required sample size for the adaptive seamless phase II/III design. We assumed the existence of a power law relationship between the overall sample size and statistical power of the final test and took into consideration the correlation between the early and final outcomes. The required sample size is defined as the minimum sample size that provides adequate power with overall type I error rate under control. The methodology was applied to determine sample sizes in a study for a treatment that can avoid renal damage during cardiac operations while the most effective dose of the treatment is unknown and will be selected at the interim analysis.

email: zhx17@pitt.edu

A BAYESIAN ADAPTIVE BASKET TRIAL DESIGN FOR RELATED DISEASES USING HETEROGENEOUS ENDPOINTS

Matthew Austin Psioda*, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Jiawei Xu, University of North Carolina, Chapel Hill
Tony Jiang, Amgen
Amy Xia, Amgen

Investigational products (IP) are being developed for specific targets and therefore may exhibit efficacy for a variety of diseases for which the target is implicated. In the oncology setting, the different diseases often correspond to different tumor histologies having a common genomic alteration (e.g., mutation). Binary endpoints are typically employed and information on treatment efficacy is borrowed across tumor histologies to increase the efficiency of the trial. Such trials are often referred to as Basket Trials. Our development is motivated by applications where the treatment effect for different diseases (i.e., baskets) will be evaluated using qualitatively different endpoints. Our adaptive design evaluates the effectiveness of the IP compared to disease-specific controls with a goal of establishing superiority of the IP in each disease while borrowing information. The approach uses Bayesian Model Averaging (BMA) and performs inference by averaging over combinations of optimistic and pessimistic priors for the parameters in each basket's data distribution.

email: matt_psioda@unc.edu

ADAPTIVELY MONITORING CLINICAL TRIALS WITH SECOND-GENERATION P-VALUES

Jonathan J. Chipman*, Vanderbilt University
Robert A. Greevy, Jr., Vanderbilt University
Lindsay Mayberry, Vanderbilt University
Jeffrey D. Blume, Vanderbilt University

The FDA is committed to "facilitate the advancement and use of complex adaptive, Bayesian, and other novel clinical trial designs". We introduce a novel, sequential monitoring design based on the Second-generation p-value (SGPV; Blume, 2018), which indicates when the data are compatible with the alternative hypothesis, the null hypothesis, or when inconclusive. This requires pre-specifyied trivial and clinically meaningful effect sizes. False discovery rate reduces by ignoring statistically significant results for trivial effect sizes. SGPVs are easy to implement and outperform traditional approaches based on adjusting the p-value for multiple comparisons or looks. The trial halts when the data support either convincingly superior or uninteresting clinical results and, to reduce bias, is affirmed by a subsequent validation monitoring examination. In extensive simulations and the currently active clinical trial REACH, we compare our method's performance in terms of error rates, bias, and average stopping times to interval null Bayesian Adaptive designs (Kruschke, 2013) and provide recommendations on implementing the stopping validation monitoring step.

email: jonathan.chipman@vanderbilt.edu

83. BAYESIAN APPROACHES TO SURVEYS AND SPACIO-TEMPORAL MODELING

ESTIMATING OF PROSTATE CANCER INCIDENCE RATES USING SERIALLY CORRELATED GENERALIZED MULTIVARIATE MODELS

Manoj Pathak*, Murray State University
Jane L. Meza, University of Nebraska Medical Center
Kent M. Eskridge, University of Nebraska, Lincoln

In recent decades, disease mapping has drawn much attention worldwide. Due to the availability of Markov Chain Monte Carlo (MCMC) algorithms, fully Bayesian analysis of complex multistage data has been increasingly popular in the study of geographically and temporally referenced data. Disease incidence or mortality data are measured longitudinally on the same geographic units. Multiple measures over time in the same region may be serially correlated by sharing similar risk factors or socio-economic status of the background population. The evolution of disease may be different at the different time periods. Due to the neighborhood structure of the spatial data, disease outcomes in the geographic unit at the time may also depend on the adjacent units at the time. In this study, we develop a serially correlated generalized multivariate conditional autoregressive model (SCGMCAR) with different propriety parameters for different time periods. The proposed methods are implemented to analyze spatially referenced longitudinal data at the small area. Analysis of Nebraska Cancer Registry data shows that introducing different propriety parameters provides a better fit.

email: mpathak@murraystate.edu

BAYESIAN HIERARCHICAL MODELS FOR VOXEL-WISE CLASSIFICATION OF PROSTATE CANCER ACCOUNTING FOR SPATIAL CORRELATION AND BETWEEN-PATIENT HETEROGENEITY IN THE MULTI-PARAMETRIC MRI DATA

Jin Jin*, University of Minnesota
Lin Zhang, University of Minnesota
Ethan Leng, University of Minnesota
Gregory J. Metzger, University of Minnesota
Joseph S. Koopmeiners, University of Minnesota

Multi-parametric MRI (mpMRI) plays an increasingly important role in the prostate cancer diagnosis. There are certain mpMRI features including substantial spatial correlation between voxels and between-patient differences in the mpMRI parameters, which have not been fully explored in the literature but can potentially improve cancer detection if leveraged appropriately. This paper proposes Bayesian models to improve classification by modeling the spatial correlation and patient heterogeneity in mpMRI. Properly modeling the spatial correlation is challenging due to the high-resolution of mpMRI. We propose three approaches using Nearest Neighbor Gaussian Process (NNGP), low-rank approximation and a conditional autoregressive (CAR) model. Simulation results showed that the proposed models achieved substantially improved classification accuracy. Real data application showed that classification was improved by spatial modeling using the NNGP and low-rank models but not the CAR model, and modeling patient heterogeneity did not add further improvement. Among the spatial modeling approaches, NNGP is recommended with robust classification accuracy and high computational efficiency.

email: jinxx493@umn.edu

A BAYESIAN SHAPE INVARIANT GROWTH CURVE MODEL FOR LONGITUDINAL DATA

Mohammad Alfrad Nobel Bhuiyan*, Cincinnati Children's Hospital Medical Center
Heidi Sucharew, Cincinnati Children's Hospital Medical Center
Md Monir Hossain, Cincinnati Children's Hospital Medical Center

The longitudinal growth curve modeling is a popular area of research because of the advantages of analyzing both the within subject effect and between subject effect simultaneously. Recently, Cole et al. suggested a non-parametric model the superimposition by Translation and Rotation (SITAR) based on shape invariant model, which express individual growth curves through three subject specific parameters. To address limitations of existing studies and better characterize the relationship between stimulant medication exposure and growth, we conducted a longitudinal study of children with ADHD. Longitudinal height measurements were modeled using the SITAR model which characterizes height trajectories in terms of size, tempo, and velocity parameters. We found early age of medication start to be associated with lower size and lower tempo. We demonstrate and compare results of the SITAR model utilizing a nonlinear mixed effects model framework to our proposed Bayesian framework approach both utilizing a natural cubic spline function in evaluating the association of age at start of stimulant medication with height trajectories.

email: bhuiyama@mail.uc.edu

BAYESIAN RECORD LINKAGE UNDER LIMITED LINKING INFORMATION

Mingyang Shan*, Brown University
Kali Thomas, Brown University
Roee Gutman, Brown University

Record linkage is a statistical technique that seeks to identify individuals or entities that refer to the same unit across two or more data sources. This is a challenging task when no unique identification variable is present. Bayesian file linking procedures have been developed that primarily focus on classifying links using semi-identifying information that exist in both data files. These procedures struggle in performance when the amount of linking variables are few in number and when the variables are prone to error. An adaptation to the existing Bayesian record linkage methodology is proposed that incorporates associations between variables that exist in only one file, in addition to those shared between both files, in order to extract additional linking accuracy when the quantity or quality of identifying information is limited. Various ways to incorporate such information into the existing record linkage framework are discussed and our methodology is applied to link Meals on Wheels recipients to Medicare Enrollment records.

email: mingyang_shan@brown.edu

BAYESIAN VARIABLE SELECTION IN GROWTH MIXTURE MODEL WITH MISSING COVARIATES DATA

Zihang Lu*, University of Toronto
Wendy Lou, University of Toronto

Studies of the growth patterns of longitudinal characteristics in early life of children destined to be healthy are of vital importance to improve our understanding in the development course of the disease. In these studies, it is often of great interest to cluster individual trajectories based on repeated measurements taken over time. Growth mixture models are commonly used in such cases to identify subgroups of the trajectory patterns. Despite its importance in facilitating medical findings, little work has been done in selecting the predictors related to class membership in the context of growth mixture models. Motivated by a Canadian birth cohort study, we propose a growth mixture model with Bayesian variable selection feature for clustering longitudinal growth trajectories, and selecting important covariates that are associated with the class-membership. Our approach provides simultaneous imputation of missing mixed-types (e.g. continuous, categorical, ordinal) covariates and variables selection in the growth mixture models context, in which longitudinal growth trajectories are modelled and subjects are clustered into subgroups conditional on their covariates.

email: zihang.lu@mail.utoronto.ca

84. CAUSAL EFFECTS WITH PROPENSITY SCORES/ WEIGHTING/MATCHING

USING PROPENSITY SCORES WITH TREATMENT SELECTION BIOMARKERS

Hulya Kocyigit*, University of Georgia

Scientists are increasingly using observational studies to estimate the effects of treatments and exposures on health outcomes. One popular statistical methodology to estimate the difference between two groups is the propensity score method. In this study, we propose an illustration of steps in propensity score methodology. Firstly, we consider propensity score matching, where distance is only evaluated in terms of the difference in the estimated propensity score and then apply trimming to propensity score.Rubin and Imbens propose that advantage is the internal validity may be improved because estimators for casual effects in trimmed sample are likely to be more credible and accurate than estimators for causal effects in the original, full sample. The propensity score method is next illustrated using subsampling on propensity score which is applied within subgroups.Up to this point, our goal was to understand how to improve balance in covariate distributions using three propensity score approaches with cancer data.Lastly, we want to evaluate a predictive biomarker for chemotherapy treatment in cancer data and evaluate how treatment efficacy varies among subsets of patients.

email: hk20902@uga.edu

BUILDING REPRESENTATIVE MATCHED SAMPLES WITH MULTI-VALUED TREATMENTS

Magdalena Bennett*, Columbia University
Juan Pablo Vielma, Massachusetts Institute of Technology
Jose R. Zubizarreta, Harvard University

In this paper, we estimate the effect that the level of exposure to the 2010 Chilean earthquake had on college admission exams. Applying the idea of template matching of Silber et al. (2014), and leveraging recent developments in optimization, these new methods allow us: (i) to handle multi-valued treatments without estimation of the generalized propensity score; (ii) to build self-weighted matched samples that are representative of a target population; and (ii) to work with much larger data sets than other similar methods. Concretely, we propose a projective formulation for matching with distributional covariate balance. We formally show that this formulation is more effective than related formulations, as it is smaller in terms of the number of decision variables, but equally strong from the standpoint of its linear programming relaxation. We implement this formulation in the package designmatch for R. The results from our case study are striking: while increasing levels of exposure to the earthquake have a negative impact on school attendance, there is no effect on college admission test scores.

email: mb3863@columbia.edu

TRIPLET MATCHING FOR ESTIMATING CAUSAL EFFECTS WITH THREE TREATMENT ARMS: A COMPARATIVE STUDY OF MORTALITY BY TRAUMA CENTER LEVEL

Giovanni Nattino*, The Ohio State University
Bo Lu, The Ohio State University
Junxin Shi, The Research Institute of Nationwide Children's Hospital
Stanley Lemeshow, The Ohio State University
Henry Xiang, The Research Institute of Nationwide Children's Hospital

Comparing outcomes across trauma center levels is vital in evaluating trauma care. Propensity score matching is a robust method to infer causal relationships in observational studies with two treatment arms. Few studies, however, have used matching designs with more than two groups, due to the complexity of matching algorithms. We fill the gap by developing a conditional three-way matching algorithm that outperforms the nearest neighbor algorithm. We illustrate an implementation of Rosenbaum's framework of evidence factors for binary outcomes, which can be used to conduct an outcome analysis and a sensitivity analysis for hidden bias on three-group matched designs. We apply our method to the Nationwide Emergency Department Sample data to compare mortality among non-trauma, level I and level II trauma centers. We find strong evidence that the admission to a trauma center has a beneficial effect on the outcome. However, the difference in mortality between level I and level II centers is not significant. The sensitivity analysis shows that unmeasured confounders moderately associated with the type of care received may change the result qualitatively.

email: nattino.1@osu.edu

A NOVEL PROPENSITY SCORE FRAMEWORK FOR A CONTINUOUS TREATMENT USING THE CUMULATIVE DISTRIBUTION FUNCTION

Derek W. Brown*, University of Texas Health Science Center at Houston
Thomas J. Greene, GlaxoSmithKline
Michael D. Swartz, University of Texas Health Science Center at Houston
Anna V. Wilkinson, University of Texas Health Science Center at Houston
Stacia M. DeSantis, University of Texas Health Science Center at Houston

Current propensity score methods rely on weighting in order to produce causal estimates in observational studies with continuous treatments. Weighting methods can result in worse covariate balance than if no adjustment had been made. Furthermore, weighting methods are not always stable and may produce unreliable estimates, due to extreme weights. These issues motivate the development of novel propensity score stratification techniques to be used with continuous treatments. Specifically, the generalized propensity score cumulative distribution function (GPS-CDF) and the nonparametric GPS-CDF (npGPS-CDF) approaches are used to stratify subjects based on pretreatment covariates to produce causal estimates. Simulation results show superiority of these new stratification methods based on the CDF when compared to standard weighting techniques. The proposed methods are applied to the Mexican-American Tobacco use in Children study to quantify the relationship between exposure to smoking imagery and smoking behavior among Mexican-American adolescents. The promising results presented here provide investigators with new options for implementing continuous treatment propensity scoring.

email: derek.brown@uth.tmc.edu

MINIMAL DISPERSION APPROXIMATELY BALANCING WEIGHTS: ASYMPTOTIC PROPERTIES AND PRACTICAL CONSIDERATIONS

Jose Zubizarreta*, Harvard University
Yixin Wang, Columbia University

Weighting methods are widely used to adjust for covariates in observational studies, sample surveys, and regression settings. In this paper, we study a class of recent weighting methods that find the weights of minimum dispersion that approximately balance the covariates. We call these weights minimal weights and study them under a common optimization framework. From a theoretical standpoint, we characterize the asymptotic properties of minimal weights and show that minimal weights are consistent estimates of the true inverse probability weights. Also, the resulting weighting estimator is consistent, asymptotically normal, and semiparametrically efficient. From a practical standpoint, we present a finite sample oracle inequality that bounds the loss incurred by balancing more functions of the covariates than needed. It shows that minimal weights implicitly bound the number of active covariate balance constraints. We also provide a tuning algorithm for choosing the degree of approximate balance. We conclude with four empirical studies that suggest approximate balance is preferable to exact balance, especially when there is limited overlap in covariate distributions.

email: yixin.wang@columbia.edu

CONDUCTING MENDELIAN RANDOMIZATION ANALYSIS ON SUMMARY DATA UNDER CASE-CONTROL STUDIES

Han Zhang*, National Cancer Institute, National Institutes of Health
Lu Deng, National Cancer Institute, National Institutes of Health
Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Kai Yu, National Cancer Institute, National Institutes of Health

Mendelian randomization analysis is used to estimate unconfounded causal effects of an exposure. The application of such methods become popular in epidemiological studies since they are extended to summary data of genome-wide association studies. Early methods focused on quantitative outcomes and later methods on prospective binary outcomes. The latter methods have been applied to case-control data with little theoretical justification. As a result, those methods give highly biased estimate for casual effect and inaccurate estimate of standard error, resulting in low coverage probability of Wald-type confidence intervals. We give assumptions that justify a Mendelian randomization analysis for case-control data and propose a novel estimate for causal effect based on a hybrid empirical likelihood method. The use of the hybrid strategy reduces the parameter space in numerical optimization, so that our method can be applied to hundreds of instruments. Compared to existing methods, our estimate is less biased, and the confidence interval derived from a Lagrange multiplier test has coverage nearest the nominal level even if instruments are weak.

email: zhangh12@mail.nih.gov

85. META-ANALYSIS

BAYESIAN NETWORK META-ANALYSIS OF TREATMENT TOXICITIES

Aniko Szabo*, Medical College of Wisconsin
Binod Dhakal, Medical College of Wisconsin

Multiple methods for network meta-analysis of binary outcomes have been proposed, however they are not directly applicable to evaluating treatment toxicities as outcomes due to several unique features. Reporting of toxicities varies between publications, and often only the most common events are reported, leading to informative missingness. The number of different toxicity types is usually large, with correlations expected due to relationship to the same organ system. Finally, individual treatments typically have characteristic side-effect profiles, and current approaches do not consider the relationship between the outcomes of combination treatments and their individual components. In this work we develop a Bayesian arm-based meta-analysis model for multivariate analysis of toxicities incorporating left-censoring and treatment-component information. The proposed methodology is applied to comparing toxicities of treatments for relapsed or refractory multiple myeloma.

email: aszabo@mcw.edu

ADAPTIVE WEIGHTING METHODS FOR IDENTIFYING CONCORDANT DIFFERENTIALLY EXPRESSED GENES IN OMICS META-ANALYSIS

Chien-Wei Lin*, Medical College of Wisconsin
George C. Tseng, University of Pittsburgh

With the rapid advances and prevalence of high throughput genomic technologies, integrating information from multiple relevant genomic studies has brought new insights into biomedical research. Meta-analysis has been widely used in this manner and combining p-values from multiple studies for differentially expressed (DE) gene detection has a long history in statistical science. Among different p-values-based methods, adaptively weighted Fisher's (AW-Fisher's) method can select gene-specific heterogeneous DE signal across studies. However, when combining two-sided p-values for binary outcomes, the exiting AW-Fisher's method does not have the advantage of filtering discordant biomarkers such that DE genes with discordant DE direction can often be identified. In this work, we propose an idea inspired by Owen's one-sided correction method to AW-Fisher's method. The results of simulations and applications in real data showed that the methods with one-sided correction are helpful to guarantee identification of DE genes with concordant DE direction without losing statistical power. An R package will be provided.

email: chlin@mcw.edu

A BAYESIAN HIERARCHICAL MODEL ESTIMATING CACE IN META-ANALYSIS OF RANDOMIZED CLINICAL TRIALS WITH NONCOMPLIANCE

Jincheng Zhou*, University of Minnesota
Haitao Chu, University of Minnesota
James S. Hodges, University of Minnesota
M. Fareed K. Suri, University of Minnesota

Noncompliance to assigned treatment is a common challenge in the analysis and interpretation of randomized clinical trials. The complier average causal effect (CACE) approach provides a useful tool for addressing noncompliance, where CACE is defined as the average difference in potential outcomes for the response in a subpopulation of subjects who comply with their assigned treatments. In this article, we present a Bayesian hierarchical model to estimate the CACE in a meta-analysis of randomized clinical trials where compliance may be heterogeneous between studies. Between-study heterogeneity is taken into account with study-specific random effects. The results are illustrated by a re-analysis of a meta-analysis comparing epidural analgesia versus no or other analgesia in labor on the outcome of cesarean section, where noncompliance varied between studies. Finally, we present comprehensive simulations evaluating the performance of the proposed approach, and illustrate the importance of including appropriate random effects and the impact of over- and under-fitting.

email: jzhou@umn.edu

QUANTIFYING THE EVIDENCE OF SELECTIVE PUBLISHING IN NETWORK META-ANALYSIS: AN EM ALGORITHM-BASED APPROACH

Arielle K. Marks-Anglin*, University of Pennsylvania
Jin Piao, University of Southern California
Jing Ning, University of Texas MD Anderson Cancer Center
Chongliang Luo, University of Pennsylvania
Yong Chen, University of Pennsylvania

Evidence-based medicine aims to optimize healthcare decision-making by leveraging results from well-conducted research, often relying on evidence synthesis from systematic reviews and meta-analyses. Network meta-analysis (NMA) is particularly useful for drawing new comparisons and ranking multiple interventions. However, recommendations can be misled if published results are a selective sample of what has been collected by trialists. Studies have shown that trials with statistically significant findings are more likely to be published than those with non-significant results. Unfortunately, very few methods properly quantify and adjust for publication bias, and the numerous parameters involved in modeling NMAs pose unique computational challenges, such that sensitivity analysis is used in practice. Motivated by this important methodological gap, we developed a novel EM algorithm for quantifying and correcting for publication bias in the network setting. We validate the method through simulation studies and calibrate against a `gold standard' analysis of published and unpublished trials from a recent NMA comparing antidepressants for major depressive disorder in adults.

email: anglinar@pennmedicine.upenn.edu

HIGH RESOLUTION FINE-MAPPING OF 406 SMOKING/DRINKING LOCI VIA A NOVEL METHOD THAT SYNTHESIZES THE ANALYSIS OF EXOME-WIDE AND GENOME-WIDE ASSOCIATION STATISTICS

Yu Jiang*, Penn State College of Medicine
Dajiang Liu, Penn State College of Medicine

Recently, our GSCAN consortium meta-analysis identified 406 loci associated with alcohol and nicotine use, using >1 million individuals. Functional dissection of these loci can lead to considerable advancement for addiction genetics. To proceed, we aggregated summary statistics of exome-array data and performed fine-mapping analysis. Not all contributing studies have both exome array and GWAS data. The association statistics for many variants were thus missing from some contributing studies. This missingness may skew the correlation between marginal statistics, and lead to the incorrect determination of causal variants using existing fine-mapping methods. We developed a novel method called partial correlation-based scores statistic (PCBS), which allows the correct estimation of joint effects when the contributed summary statistics contain missing data. We further extended this method to incorporate functional genomic data from ENCODE, Roadmap and GTEx. The PCBS-based methods are particularly useful for the next phase of fine-mapping studies, where GWAS and sequence data (e.g. TOPMed) that are not measured on all study subjects are synthesized.

email: ybj5037@psu.edu

BAYESIAN NETWORK META-REGRESSION FOR ORDINAL OUTCOMES INCORPORATING HIGH-DIMENSIONAL RANDOM EFFECTS

Yeongjin Gwon*, University of Nebraska Medical Center
Ming-Hui Chen, University of Connecticut
May Mo, Amgen Inc.
Jiang Xun, Amgen Inc.
Amy Xia, Amgen Inc.
Joseph Ibrahim, University of North Carolina, Chapel Hill

In this paper, we propose an arm-based network meta-regression approach for modeling ordinal outcomes under logit link. Specifically, we develop regression model based on aggregate treatment-level covariates for the underlying cut-off points as well as for the variances of random effects to capture heterogeneity across trials. To incorporate high-dimensional random effects, we utilizes a data augmentation strategy using Polya-Gamma latent variables. We then develop an efficient computational algorithm and it allows for more flexible modeling strategy for the variances of random effects. We further develop Bayesian model comparison measures to assess of goodness-of-fit and the determination of outlying trials. A case study demonstrating the usefulness of the proposed methodology is carried out to assess the relative effectiveness of different treatment options in treating Crohn's disease.

email: yeongjin.gwon@unmc.edu

MULTIVARIATE META-ANALYSIS OF RANDOMIZED CONTROLLED TRIALS WITH THE DIFFERENCE IN RESTRICTED MEAN SURVIVAL TIMES

Isabelle R. Weir*, Boston University School of Public Health
Lu Tian, Stanford University
Ludovic Trinquart, Boston University School of Public Health

The difference in restricted mean survival times (RMSTD) offers an absolute sure of the treatment effect on the time scale. Computation of the RMSTD relies on the choice of a time horizon t*. In a meta-analysis of randomized controlled trials (RCTs), varying follow-up durations may lead to the exclusion of RCTs with follow-up shorter than t*. We introduce a multivariate meta-analysis model for RMSTD at multiple time horizons. We derived the within-trial covariance for the RMSTD at multiple time points. The model enables the synthesis of all observed data by borrowing strength from multiple time points. In a simulation study, we compared the statistical performance of the proposed method to that of two univariate meta-analysis models, based on observed data and based on predictions from fexible parametric models. Our multivariate model yields smaller mean squared error over univariate methods. We illustrate the approach using 5 RCTs of transcatheter aortic valve implantation in patients with aortic stenosis at 12, 24, and 36 months.

email: iweir@bu.edu

86. IMAGING APPLICATIONS AND TESTING

SEMIPARAMETRIC MODELING OF TIME-VARYING ACTIVATION AND CONNECTIVITY IN TASK-BASED fMRI DATA

Jun Young Park*, University of Minnesota
Joerg Polzehl, Weierstrass Institute for Applied Analysis and Stochastics
Snigdhansu Chatterjee, University of Minnesota
André Brechmann, Leibniz-Institute for Neurobiology
Mark Fiecas, University of Minnesota

In fMRI, there is a rise in evidence that the dynamic functional connectivity (dFC) provides additional information on brain networks not captured by static connectivity. While there have been many statistical models for dFC whenever the study participants are at rest, there remains a gap in the literature on how to model dFC whenever the study participants are undergoing an experimental task designed to probe at a cognitive process of interest. We propose a method to estimate the dFC between two regions of interest in task-based fMRI time series data where the activation effects are also allowed to be varying over time. Our method uses penalized spline to model both time-varying activation effects and connectivity and uses the bootstrap for statistical inference. We validate our approach using simulations and show that ignoring time-varying activation effects would lead to poor estimation of dFC. Our approach effectively estimates the true activation effects and connectivity, while being robust whenever the activation and connectivity of the brain are static. We apply our method to a fMRI learning experiment and show its relations to the behavioral data.

email: park1131@umn.edu

ON PREDICTABILITY AND REPRODUCIBILITY OF INDIVIDUAL FUNCTIONAL CONNECTIVITY NETWORKS FROM CLINICAL CHARACTERISTICS

Emily L. Morris*, University of Michigan
Jian Kang, University of Michigan

In recent years, understanding functional brain connectivity has become increasingly prominent and meaningful, both clinically and scientifically. Many statistical methods, such as graphical models and network analysis, have been adopted to construct functional connectivity networks (FCNs) for single subjects using resting state fMRI data. It is of great interest to understand the consistency and discrepancy of the constructed FCNs across multiple individuals. Here, we focus on studying the association between FCNs and clinical characteristics such as neurological symptoms and diagnoses. Utilizing the state-of-the-art machine learning algorithms, we propose a method to examine predictability of FCNs from clinical characteristics. Our methods can identify the important clinical characteristics that are predictive of the whole brain network or some subnetworks. Using persistent homology, our methods also can measure the reproducibility of constructing multiple individual FCNs at different spatial resolutions. We illustrate our methods on the analysis of fMRI data in the Philadelphia Neurodevelopmental Cohort (PNC) study and obtain some clinically meaningful results.

email: emorrisl@umich.edu

ON STATISTICAL TESTS OF FUNCTIONAL CONNECTOME FINGERPRINTING

Zeyi Wang*, Johns Hopkins Bloomberg School of Public Health
Haris Sair, Johns Hopkins University School of Medicine
Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health
Martin Lindquist, Johns Hopkins Bloomberg School of Public Health
Bennett A. Landman, Vanderbilt University
Susan Resnick, National Institute on Aging, National Institutes of Health
Joshua T. Vogelstein, Johns Hopkins University School of Medicine
Brian Caffo, Johns Hopkins Bloomberg School of Public Health

Fingerprinting of functional connectomes is an increasingly standard measure of reproducibility in functional magnetic resonance imaging (fMRI) connectomics. In such studies, one attempts to match a subject's first session image with their second, in a blinded fashion, in a group of subjects measured twice. The number or percentage of correct matches is usually reported as a statistic. In this manuscript, we investigate the statistical tests of matching based on exchangeability assumption. We show that a nearly universal Poisson (1) approximation applies for different matching schemes. We theoretically investigate the permutation tests and explore the issue that the test is overly sensitive to uninteresting directions in the alternative hypothesis, such as clustering due to familial status or demographics. We perform a numerical study on two fMRI resting state datasets, the Human Connectome Project (HCP), which has technical replications of long scans and includes twins, and the Baltimore Longitudinal Study of Aging (BLSA), which has more typical length scans in a longitudinal study. Finally, a study of single regional connections is performed on the HCP data.

email: zwang107@gmail.com

A COMPARISON OF MACHINE LEARNING ALGORITHMS FOR PREDICTING AGE FROM MULTIMODAL NEUROIMAGING DATA

Joanne Beer*, University of Pennsylvania
Helmet Karim, University of Pittsburgh
Dana Tudorascu, University of Pittsburgh
Howard Aizenstein, University of Pittsburgh
Stewart Anderson, University of Pittsburgh
Robert Krafty, University of Pittsburgh

While much prior work has focused on classification of individuals into diagnostic groups based on neuroimaging data, relatively less attention has been paid to predicting continuous variables. Recently there has been increasing interest in predicting brain age from neuroimaging studies. The difference between an individual's actual and predicted age can serve as a biomarker indicating their degree of deviation from a typical healthy brain aging trajectory. In this work, we compare various algorithms for predicting age from structural and functional neuroimaging data in a sample of cognitively normal older adults. Methods include structured sparse penalized linear regression, random forest, support vector, and Gaussian process regression. We also investigate different feature screening methods based on linear or nonlinear association with age, and using univariable or multivariable approaches. We discuss the accuracy and interpretability of different methods as well as their strengths and weaknesses in comparison to currently popular deep learning approaches.

email: joannecbeer@gmail.com

A STATISTICAL MODEL FOR STOCHASTIC RADIOGRAPHIC LUNG CHANGE FOLLOWING RADIOTHERAPY OF LUNG CANCER

Nitai Das Mukhopadhyay*, Virginia Commonwealth University
Viviana A. Rodriguez, Virginia Commonwealth University

Radiotherapy is the standard treatment for inoperable patients with non-small cell lung tumor (NSCLC). Healthy tissue near the tumor is exposed to external beam RT causing radiographic radiation-induced lung damage (RILD). In this study, we aim to build a statistical model of RILD over time by assembling smaller models that describe location-specific changes. Five CT-scans of one NSCLC patient were obtained during 21 months following RT. CT-scans are 3D images with voxels of approximately 2-3 mm in superior-inferior direction, and <1 mm in the other two directions. Integrity of the voxel tissue across time cannot be assured, hence we attempt to combine the voxels into larger subvolumes, called patches. we propose grouping the voxels as square prisms patches consisting of p voxels along each direction. The extent of RILD is measured through three-threshold based ordinal categories of radiographic injury, namely, dense, hazy and none. Each patch is represented by a vector in the form of compositional data. We will present an approach of determining the size of the patch as well as a statistical model that captures the complexity of the data structure.

email: nitai.mukhopadhyay@vcuhealth.org

AN INTER-FEATURE CORRELATION GUIDED CLASSIFIER FOR ALZHEIMER'S DISEASE PREDICTION

Yanming Li*, University of Michigan

We propose an ultrahigh-dimensional feature screening and classification method for predicting Alzheimer's disease (AD) status based on individual positron emission tomography (PET) brain imaging scans. Leveraging the spatial inter-feature correlations, the proposed method can effectively and efficiently select the voxel PET imaging predictors that are predictive for Alzheimer's disease status, especially for those predictors with marginally weak effects. we show that the proposed classifier recovers informative features with probability tending to one and can asymptotically achieve a zero misclassification rate. We evaluate the finite sample performance of the method via simulations and apply this method to ADNI PET images to classify individual AD disease status.

email: liyanmin@umich.edu

87. METHODOLOGICAL CHALLENGES AND OPPORTUNITIES IN MENTAL HEALTH RESEARCH

INTEGRATIVE LEARNING TO COMBINE INDIVIDUALIZED TREATMENT RULES FROM MULTIPLE RANDOMIZED TRIALS

Yuanjia Wang*, Columbia University

Implementing individualized treatment rules (ITRs) that adapt to patient's characteristics and intermediate responses holds promise to improve treatment response of mental disorders. However, several barriers, in particular, lack of power to detect treatment modifiers as tailoring variables and lack of generalizability or

reproducibility of ITRs derived from a single study, pose significant challenges to clinical practice. In this work, we propose a novel integrative learning method to combine evidence from multiple clinical trials to yield an integrative ITR that improves both precision and reproducibility. Since subject-specific covariates available in each trial may differ, ITRs learned from each study can depend on a different resolution of patient-specific characteristics. Our method does not require all studies to use the same set of covariates, and thus allows study-specific ITRs to be transferable across studies. We apply the developed method to multiple clinical trials of major depressive disorder and other co-morbid mental disorders.

email: yw2016@cumc.columbia.edu

MIXED-EFFECTS MODELING TO COMPARE DYNAMIC TREATMENT REGIMENS WITH SMART DATA

Brook Luers*, University of Michigan
Min Qian, Columbia University
Inbal Nahum-Shani, University of Michigan
Connie Kasari, Unviersity of California, Los Angeles
Daniel Almirall, University of Michigan

A dynamic treatment regimen (DTR) is a sequence of decision rules, each of which recommends a treatment based on a patient's past and current health status, including response to prior treatment. Sequential Multiple Assignment Randomized Trials (SMART) are innovative, multi-stage trial designs that yield data specifically for building effective DTRs. Most SMARTs include a set of DTRs that are embedded within the trial. An important primary aim in a SMART is the comparison of the embedded DTRs based on change in a primary repeated measures outcome. This manuscript focuses on the development of a mixed-effects model for the comparison of embedded DTRs on a repeated-measures outcome. The methodology is illustrated using data from two SMARTs, one in addictions and another in autism.

email: luers@umich.edu

HANDLING MISSING CLINICAL AND MULTIMODAL IMAGING DATA IN INTEGRATIVE ANALYSIS WITH APPLICATIONS TO MENTAL HEALTH RESEARCH

Adam Ciarleglio*, The George Washington University
Eva Petkova, New York University

In mental health research, the number of studies that include multimodal neuroimaging is growing. Often, the goal is to integrate both the clinical and imaging data to address a specific research question. Functional data analytic tools for analyzing such data can perform well, but these methods assume complete data. In practice, some proportion of the data may be missing. We present approaches for imputation of missing scalar and functional data when the goal is to fit a scalar-on-function regression model for the purpose of either (1) estimating the association between a scalar outcome and a scalar or functional predictor or (2) developing a predictive model. We present results from a simulation study showing the performance of various imputation approaches with respect to fidelity to the observed data, estimation of the parameters of interest, and prediction. The proposed approaches are illustrated using data from a placebo-controlled clinical trial assessing the effect of SSRI in subjects with major depressive disorder.

email: aciarleglio@gwu.edu

DESIGN AND ANALYTIC TOOLS FOR PERSONALIZING HEALTHCARE EXPERIMENTS

Christopher H. Schmid*, Brown University

Single-case experimental designs can be used to create personalized protocols to make personalized treatment decisions. The N-of-1 trial uses a multi-crossover randomized design to measure individual treatment efficacy. Combining trials in a multilevel structure enables assessment of average treatment effects in populations and treatment effect heterogeneity in subgroups. I discuss some completed and ongoing N-of-1 studies using mobile device applications with server-driven statistical analytics to return results to individuals. Issues that arise include defining treatments and sequences of treatments, synthesizing treatment networks, incorporating patient-specific prior information, automating the choice of appropriate statistical models and assessment of model assumptions, and automating graphical displays and text to facilitate appropriate interpretation by non-technical users. Development of smart tools that solve these problems could help to transform health care research by expanding the settings in which it is carried out and making findings directly applicable to and interpretable by individual trial participants.

email: christopher_schmid@brown.edu

88. NOVEL APPROACHES FOR GROUP TESTING FOR ESTIMATION IN BIOSTATISTICS

MISCLASSIFIED GROUP TESTED CURRENT STATUS DATA

Nicholas P. Jewell*, London School of Hygiene & Tropical Medicine
Lucia Petitio, Harvard T.H. Chan School of Public Health

Group testing reduces costs for prevalence estimation based on screening k groups of n individuals. For low prevalence, and misclassified screening tests, more precision can be obtained than from testing all n samples separately. When the binary response indicates that a time to incidence variable T is less than a screening time C, i.e. current status data, one can consistently estimate the distribution function F of T nonparametrically for individually tested results, at least at some points in the support. We consider nonparametric estimation of F for group tested current status data where a group tests positive if and only if any individual unobserved T in the group is less than its corresponding observed C. We consider cost savings and its relationship to precision over the support of F, and investigate misclassification of the pooled tests with applications where interest focuses on the age at incidence distribution rather than prevalence.

email: jewell@berkeley.edu

GENERALIZED ADDITIVE REGRESSION FOR GROUP TESTING DATA

Joshua M. Tebbs*, University of South Carolina
Yan Liu, University of Nevada, Reno
Christopher S. McMahan, Clemson University
Chris R. Bilder, University of Nebraska, Lincoln

In screening applications involving low-prevalence diseases, pooling specimens through group testing can be far more cost effective than testing specimens individually. Estimation is a common goal in such applications and typically involves modeling the probability of disease as a function of available covariates. In recent years, several authors have developed regression methods to accommodate the complex structure of group testing data but often under the assumption that covariate effects are linear. Although linearity is a reasonable assumption in some applications, it can lead to model misspecification and biased inference in others. To offer a more flexible framework, we propose a Bayesian generalized additive regression approach to model the individual-level probability of disease with potentially misclassified group testing data. Our approach can be used to analyze data arising from any group testing protocol with the goal of estimating multiple unknown smooth functions of covariates, standard linear effects for other covariates, and assay accuracy probabilities. We illustrate the methods in this article using group testing surveillance data on chlamydia infection in Iowa.

email: tebbs@stat.sc.edu

GROUPING METHODS FOR ESTIMATING THE PREVALENCES OF RARE TRAITS FROM COMPLEX SURVEY DATA THAT PRESERVE CONFIDENTIALITY OF RESPONDENTS

Noorie Hyun*, Medical College of Wisconsin
Joseph L. Gastwirth, The George Washington University
Barry I. Graubard, National Cancer Institute, National Institutes of Health

This paper extends the methodology of 1-stage group testing to surveys with sample weighted complex multistage-cluster designs. Sample weighted generalized estimating equations are used to estimate the prevalences of categorical traits while accounting for the error rates inherent in the tests. Two difficulties arise when using group testing in complex samples: (1) How does one weight the results of the test on each group as the sample weights will differ among observations in the same group; (2) How does one form groups that will allow accurate estimation of the standard errors of prevalence estimates under multistage-cluster sampling allowing for intracluster correlation of the test results. We study 5 different grouping methods to address the weighting and cluster sampling aspects of complex designed samples. Finite sample properties of the estimators of prevalences and variances for these grouping methods are studied using simulations. National Health and Nutrition Examination Survey data are used to illustrate the methods.

email: nhyun@mcw.edu

NONPARAMETRIC ESTIMATION OF A CONTINUOUS DISTRIBUTION FOLLOWING GROUP TESTING

Aiyi Liu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Wei Zhang, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Qizhai Li, Chinese Academy of Sciences
Paul S. Albert, National Cancer Institute, National Institutes of Health

Group testing strategy has been frequently used in estimating the prevalence of rare diseases, in addition to its wide application for screening for the diseases. This article concerns a secondary analysis problem of estimating a continuous distribution upon the completion of group testing. The context is that in a study employing group testing to estimate the prevalence of a disease, data on a continuous biomarker are also collected corresponding to each individual study participant, and we are interested in estimating the distribution of the biomarker given the disease status. In this paper we construct nonparametric estimation of the distribution and obtain its asymptotic properties. We evaluate the performance of the distribution estimator under various design considerations concerning group sizes and classification errors. The method is exemplified with data from the National Health and Nutrition Examination Survey (NHANES) study to estimate the prevalence of chlamydia in urine samples and the distribution of blood monocyte percent in serum samples.

email: liua@mail.nih.gov

89. ADAPTIVE AND BAYESIAN ADAPTIVE DESIGN IN BIOEQUIVALENCE AND BIOSIMILAR STUDIES

OPTIMAL ADAPTIVE SEQUENTIAL DESIGNS FOR CROSSOVER BIOEQUIVALENCE STUDIES

Donald J. Schuirmann*, U.S. Food and Drug Administration

The Product Quality Research Institute (PQRI) Sequential Design Working Group has considered two-stage sequential bioequivalence (BE) designs that permit stopping at the end of stage 1 (with sufficient evidence to conclude BE), as with a group sequential design, and reestimating the sample size for stage 2 (if we have not stopped at stage 1), based on the estimated variance from stage 1, as with an "internal pilot" design. A goal was to follow the principle of controlling the overall type-I error rate for the study, as called for, for example, by the E9 Guidance (1998). We were able to find idealized designs that achieved these goals (Potvin et al. 2008, Montague et al. 2011.) More recently our group has sought to characterize designs that are more realistic (e.g. allowing an upper limit on total sample size, allowing stopping at stage 1 for futility, etc.) and that are optimal in terms of reducing expected sample size.

email: donald.schuirmann@fda.hhs.gov

A BAYESIAN ADAPTIVE DESIGN FOR BIOSIMILAR CLINICAL TRIALS USING CALIBRATED POWER PRIOR

Ying Yuan*, University of Texas MD Anderson Cancer Center

We propose a Bayesian adaptive design for two-arm randomized trials to evaluate biosimilar products. To take advantage of the abundant historical data on the efficacy of the reference product that is typically available at the time a biosimilar product is developed, we propose the calibrated power prior, which allows our design to adaptively borrow information from the historical data according to the congruence between the historical data and the new data collected from the current trial. We propose a new measure, the Bayesian biosimilarity index, to measure the similarity between the biosimilar and the reference product. During the trial, we evaluate the Bayesian biosimilarity index in a group sequential fashion based on the accumulating interim data, and stop the trial early once there is enough information to conclude or reject the similarity. Extensive simulation studies show that the proposed design has higher power than traditional designs. We applied the proposed design to a biosimilar trial for treating rheumatoid arthritis.

email: yyuan@mdanderson.org

SEQUENTIAL BIOEQUIVALENCE

A. Lawrence Gould*, Merck Research Laboratories

Bioequivalence trials, including biosimilarity trials, compare the relative biovailability of different drug formulations. A 90% confidence interval for the ratio of expected pharmacologic endpoint values of the formulations that lies between specified endpoints, e.g., 0.8 - 1.25 conventionally is required for demonstrating bioequivalence. The likelihood of demonstrating bioequivalence of truly bioequivalent formulations depends on the sample size and on the variability of the pharmacologic endpoint variable. Group sequential bioequivalence testing provides a statistically valid way to accommodate initial misspecification of the variability by allowing for additional observations if bioequivalence cannot be accepted or rejected clearly with the initial set of observations. Group sequential bioequivalence designs applicable in most practical situations allow a decision to be reached with fewer observations than fixed sample designs about 60% of the time at approximately the same average cost. Providing the capability of sequential decisions modestly affects the nominal significance levels, e.g., the required confidence level may be 93-94% instead of 90%.

email: goulda@merck.com

90. METHODS TO ROBUSTLY INCORPORATE EXTERNAL DATA INTO GENETIC TESTS

ANCESTRY-MATCHED ALLELE FREQUENCY ESTIMATES

Tracy Ke*, Harvard University
Alex Bloemendal, Broad Institute
Danfeng Chen, Broad Institute
Claire Churchhouse, Broad Institute
Benjamini Neale, Broad Institute
Duncan Palmer, Broad Institute
Klea Panayidou, Carnegie Mellon University
Katherine Tashman, Broad Institute
Kathryn Roeder, Carnegie Mellon University

We are interested in using a large number of publicly available control samples to enhance GWAS study. The Universal Control Repository Network (UNICORN) introduces a model-based approach to aggregating publicly available control samples and estimating the ancestry-matched minor allele frequency for given case samples. First, a hierarchical tree structure is built by iteratively running spectral clustering on all control samples; it generates a mapping that projects any case sample to the “ancestry space” of one leaf cluster. Second, for each leaf cluster, assuming the null MAF is a smooth function in the ancestry space, model-based methods are developed for estimating the ancestry-matched MAF. The original version of UNICORN uses Bayesian spatial kriging for the analysis on each leaf cluster. We propose a newer version that uses a parametric logistic regression for the analysis on each leaf cluster. This new version not only significantly reduces the computational cost but also makes inference more accessible. We test the new UNICORN method via both simulations and real data. We also develop a theoretical framework to justify the consistency of the methods.

email: zke@fas.harvard.edu

EMPOWERING EXTERNAL MULTI-ETHNIC DATA IN MODERN, DIVERSE STUDIES

Chris Gignoux*, Colorado Center for Personalized Medicine

Recent international efforts in human genetics have focused on developing large repositories of genomic variation. While we have made many associated and trait discoveries, we know that genetic ancestry plays a far greater role in variability genome-wide, and not accounting for this population stratification can result in inflated Type I and Type II error rates. Here, we highlight the importance of modeling population structure to provide ancestry-matching in external data, using an automatable pipeline designed for both sequencing and genotyping data from the Population Architecture using Genomics and Epidemiology (PAGE) and Genome Sequencing Program (GSP) Studies. We highlight the use of this for multi-ethnic designs, and the need for ancestry matching for rare-variant and pooled-variant tests, where the inclusion of standard ancestry covariates such as PCA may not remove confounding due to population structure.

email: chris.gignoux@ucdenver.edu

INTEGRATING EXTERNAL CONTROLS TO ASSOCIATION TESTS

Seunggeun Lee*, University of Michigan

Identifying disease-associated genetic variants requires a large number of samples. One cost-effective approach to increase the power is using external controls, whose genomes have been genotyped and publicly available. However, when using external controls, possible batch effects due to the use of different sequencing platforms or genotype calling pipelines can dramatically increase type I error rates. To address this, we have developed allele frequency based single and gene or region based tests, called iECAT, which allows the integration of external controls while controlling for type I error. Our approach is based on the insight that batch effects on a given variant can be assessed by comparing odds ratio estimates using internal controls only vs. using combined control samples of internal and external controls. Recently, we have expanded iECAT to score test to adjust for covariates. We then extend our method to incorporate batch effect information across SNPs. We show that our new approach can control for type I error rates and increase power over the original iECAT.

email: leeshawn@umich.edu

ProxECAT: A CASE-CONTROL GENE REGION ASSOCIATION TEST USING ALLELE FREQUENCIES FROM PUBLIC CONTROLS

Audrey E. Hendricks*, University of Colorado Denver

Recent investments have resulted in millions of sequenced samples. Although large studies exist, most sequencing studies are much smaller consisting of hundreds to thousands of subjects. This results in low power especially for complex diseases. Large publicly available resources, such as the Genome Aggregation Database (~140K sequenced samples), could be used as controls. However, these genetic resources are often not used or not used appropriately due to the lack of suitable statistical methods. Here we present ProxECAT (Proxy External Controls Association Test), a rare-variant burden test that uses externally sequenced controls and internally sequenced cases. ProxECAT was motivated by the observation that variants that are often discarded in burden tests can be used as a proxy for how well variants within a genetic region are called. ProxECAT incorporates this information directly into the test statistic, which maintains the expected type I error and has increased power as the number of external controls increases. In addition to presenting ProxECAT, we present extensions to incorporate study level covariates such as ancestry, and multiple external samples.

email: audrey.hendricks@ucdenver.edu

91. DEVELOPING COLLABORATIVE SKILLS FOR SUCCESSFUL CAREERS IN BIOSTATISTICS AND DATA SCIENCE

PANEL DISCUSSANTS:

Lei Shen, Eli Lilly and Company
Patrick Staples, Mindstrong Health
Eric Ross, Fox Chase Cancer Center
Barret Schloerke, RStudio

92. NEW APPROACHES TO CAUSAL INFERENCE UNDER INTERFERENCE: BRINGING METHODOLOGICAL INNOVATIONS INTO PRACTICE

DESIGN AND ANALYSIS OF VACCINE STUDIES IN THE PRESENCE OF INTERFERENCE

M. Elizabeth Halloran*, Fred Hutchinson Cancer Research Center and University of Washington

Vaccination of individuals can often have an effect on whether other individuals become infected. Thus, interference is often present in vaccination programs, and vaccination of individuals in populations can have several different kinds of effects. Here we discuss different types of studies that have been and are being conducted to evaluate these different types of effects, including cluster-randomized studies, observational studies, and the use of big data sources. We present examples of each and point to new developments using networks.

email: betz@u.washington.edu

NONPARAMETRIC IDENTIFICATION OF CAUSAL INTERVENTION EFFECTS UNDER CONTAGION

Wen Wei Loh*, Ghent University
Forrest W. Crawford, Yale University

Estimating the causal effect of an intervention (e.g., vaccination) on infectious disease outcomes is difficult because outcomes may be contagious, and interventions may affect both susceptibility prior to infection and infectiousness once infected. A simple two-person household model has been influential in helping researchers conceptualize contagion, define causal estimands, and identify effects. In this paper, we significantly generalize a canonical model of contagion to define causal intervention effects in symmetric partnerships where both individuals can be treated, and either can transmit infection to the other, in continuous time. When infection times are observed, we show that these effects are nonparametrically identified under less restrictive assumptions than those typically required by mediation approaches. We outline new causal estimands for intervention effects under contagion, and show formally why randomization is not sufficient to eliminate confounding under contagion.

email: WenWei.Loh@UGent.be

PAIRWISE REGRESSION IN INFECTIOUS DISEASE EPIDEMIOLOGY WITH APPLICATIONS TO EBOLA AND CHOLERA

Eben Kenah*, The Ohio State University

Dependent happenings in infectious disease transmission data can be handled using methods from survival analysis by modeling failure times in ordered pairs of individuals. The contact interval in the pair ij is the time from the onset of infectiousness in i until infectious contact from i to j, which is a contact sufficient to infect j if he or she is still susceptible. We show how accelerated failure time models and semiparametric relative-risk regression can be adapted to simultaneously estimate covariate effects on infectiousness and susceptibility. These methods are available in an R package called transtat that is available on GitHub. To show how they can provide novel insights into infectious disease transmission, we apply them to cholera household transmission data from Bangladesh and to data from the WHO Ebola vaccination trial in Guinea.

email: kenah.1@osu.edu

NEW APPROACHES TO CAUSAL INFERENCE UNDER INTERFERENCE: BRINGING METHODOLOGICAL INNOVATIONS INTO PRACTICE

Xiaoxuan Cai*, Yale University
M. Elizabeth Halloran, Fred Hutchinson Cancer Research Center and University of Washington
Wen Wei Loh, Ghent University
Eben Kenah, The Ohio State University
Forrest W. Crawford, Yale University

Measuring the effect of infectious disease interventions (vaccinations) is a major challenge in contemporary epidemiology because the outcome of interest – infection – is transmissible between individuals. This complication means that individuals' infection outcomes may depend on the treatments and outcomes of other individuals, also known as "interference" or "spillover". In our work, we propose a general stochastic model of infectious disease transmission in continuous time, and natural definitions of causal estimands for individual-level direct and indirect vaccine effects. This framework provides the methodological foundations for practical causal inference in interconnected study populations. We develop semi-parametric statistical models and an inferential procedure for estimating vaccine effects, which involve multiple time scales and permit regression adjustment for variables associated with susceptibility to disease, and infectiousness once infected. Large-sample statistical properties are established under the theory of counting processes, and performance of the models are verified by simulations.

email: nmcaixiaoxuan@gmail.com

93. DESIGN AND ANALYSIS OF CLINICAL TRIALS

A UNIFIED APPROACH FOR FREQUENTIST AND BAYESIAN HYPOTHESIS TESTING IN TWO-ARM FIXED-SAMPLE CLINICAL TRIALS WITH BINARY OUTCOMES

Zhenning Yu*, Medical University of South Carolina
Viswanathan Ramakrishnan, Medical University of South Carolina
Caitlyn Meinzer, Medical University of South Carolina

Two opposing paradigms, analyses via frequentist or Bayesian methods, dominate the statistical literature. Most commonly, frequentist approaches have been used to design and analyze clinical trials, though Bayesian techniques are becoming increasingly popular. However, these two paradigms can generate divergent results even when analyzing the same trial data, which may harm the scientific merit of clinical trials. Therefore, it is crucial to harmonize frequentist approaches and Bayesian approaches for clinical trial results. In this paper, we propose a unified framework for one-sided frequentist and Bayesian hypothesis tests comparing two proportions in fixed-sample clinical trials. We assume the Bayesian prior distribution to be a beta distribution, and use the posterior probability of the proportional difference as the Bayesian statistic. We show that the unified framework can yield the same type I and II error probabilities for frequentist and Bayesian hypothesis testing through a numerical study.

email: yuz@musc.edu

DESIGNING TWO ARM CLINICAL TRIALS WITH HISTORICAL DATA USING BAYESCTDESIGN

Barry S. Eggleston*, RTI International
Diane J. Catellier, RTI International
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

A goal of clinical trial design is to estimate number of subjects needed. Enrolling too many or too few subjects is unethical,. In both cases it wastes funding and exposes an unnecessary number of subjects to safety risks. To design an efficient and ethical trial, one can incorporate historical information. Bayesian methods make this very easy. One method to incorporate historical information into the design is to construct a power prior based on historical control data. The power prior is equal to the likelihood of the historical control data raised to a power that ranges from 0 to 1. In this presentation, we will illustrate an R package, BayesCTDesign, that can help a clinical trialist investigate and design a two-arm clinical trial that incorporates historical data using the power prior. We will use the package to determine the benefit of including historical control data relative to a traditional two-arm trial as well as assess risk of historical and randomized control mismatch which causes bias and Type 1 error inflation.

email: btj@mebtel.net

RANDOMIZATION INFERENCE FOR A TREATMENT EFFECT IN CLUSTER RANDOMIZED TRIALS

Dustin J. Rabideau*, Harvard T. H. Chan School of Public Health
Rui Wang, Harvard T. H. Chan School of Public Health, Harvard Medical School and Harvard Pilgrim Health Care Institute

In a cluster randomized trial (CRT), groups of people rather than individuals are randomly assigned to different interventions. This clustering induces dependence between individual-level outcomes, which is typically handled by using a generalized linear mixed model or a generalized estimating equation; however, these approaches can lead to inflated type I error and a confidence interval (CI) with lower than nominal coverage when the parametric assumptions are violated or in CRTs with a small number of clusters. Randomization inference provides an attractive alternative approach that makes no distributional assumptions and does not require a large number of clusters to be valid. We propose a unified framework for randomization inference in the CRT setting and evaluate a computationally efficient algorithm to obtain CIs. This general approach results in valid tests and corresponding CIs for a treatment effect regardless of the type of outcome (e.g. binary, time-to-event) or study design (e.g. parallel, stepped wedge). We evaluate the performance of the proposed approach through simulation studies and apply it to a large HIV prevention CRT.

email: djrabideau@g.harvard.edu

IS CORRECTING FOR MULTIPLE TESTING IN A PLATFORM TRIAL NECESSARY?

Jessica R. Overbey*, Mailman School of Public Health, Columbia University and Icahn School of Medicine at Mount Sinai
Ying Kuen K. Cheung, Mailman School of Public Health, Columbia University
Emilia Bagiella, Icahn School of Medicine at Mount Sinai

Platform trials evaluating multiple treatments against a shared control group are efficient alternatives to 2-arm trials. Whether tests of each treatment against the control need adjustment for multiple testing is not well-established. We conducted simulation studies to evaluate the operating characteristics of 3-arm platform designs adjusted and unadjusted for multiple testing relative to 2 independent 2-arm trials. Closed designs where arms enroll simultaneously and open designs where the second experimental arm is introduced after the first reaches 25% enrollment were evaluated. Familywise error rates (FWER) and conditional errors rates (CER) were explored. Simulation results show that when unadjusted, platform trials yield similar FWER compared to equivalent 2-arm trials; however, CER is substantially higher. When early stopping is not allowed, CER is 18% and 12% in closed and open respectively versus 2.5% in independent trials. A Bonferroni correction reduces FWER by half that of the 2-arm framework while CER remain high at 14% and 9% for closed and open designs. Given these results we conclude that correcting for multiple testing may be unnecessary in platform trials.

email: jroverbey913@gmail.com

SEQUENTIAL EVENT RATE MONITORING IN CLINICAL TRIALS

Dong-Yun Kim*, National Heart, Lung, and Blood Institute, National Institutes of Health
Sung-Min Han, Open Source Electronic Health Record Alliance (OSEHRA

In this talk, we propose Sequential Event Rate Monitoring (SERM), a new continuous monitoring method for the event rate of time-to-event data in a clinical trial. SERM gives an early warning if the target rate is unlikely to be achieved by the end of study. Since SERM is designed to monitor the overall event rate, blindness of the trial is preserved. If necessary, the method could suggest the number of extra recruitments required for the planned number of primary events. It can also be used to estimate an extension of the follow-up time. We illustrate the method using data from a well-known Phase III clinical trial.

email: kimd10@nhlbi.nih.gov

GROUP SEQUENTIAL ENRICHMENT DESIGNS BASED ON ADAPTIVE REGRESSION OF RESPONSE AND SURVIVAL TIME ON HIGH DIMENSIONAL COVARIATES

Yeonhee Park*, Medical University of South Carolina
Suyu Liu, University of Texas MD Anderson Cancer Center
Peter Thall, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center

Precision medicine relies on the idea only a subpopulation of patients is sensitive to and benefit from a targeted agent. In practice, it often is assumed that the sensitive subpopulation is known and the agent is substantively efficacious in that subpopulation. Subsequent patient data, however, often show that these assumptions are false. We provide a Bayesian randomized group sequential enrichment design to compare an experimental treatment to a control based on survival time. Early response is used as an ancillary outcome to assist with adaptive variable selection, enrichment, and futility stopping. The design starts by enrolling patients under broad eligibility criteria. At each interim, submodels for regression of response and survival time on a possibly high dimensional covariates and treatment are fit, variable selection is used to identify a covariate subvector that characterizes sensitive patients and determines a personalized benefit index, and superiority and futility decisions are made. A simulation study shows that the proposed design accurately identifies a sensitive subpopulation and yields much higher power than a conventional group sequential design.

email: funnypyh@gmail.com

94. SEMIPARAMETRIC, NONPARAMETRIC, AND EMPIRICAL LIKELIHOOD MODELS

THE BEHRENS-FISHER PROBLEM IN GENERAL FACTORIAL DESIGNS WITH COVARIATES

Cong Cao*, University of Texas, Dallas
Frank Konietschke, University of Texas, Dallas

In many disciplines, factorial designs are widely applied to test the treatment effects involved in the trials. The ANOVA-Type Statistic (ATS) with Box-type approximation is often used in nonparametric factorial designs. When the covariates are present, the factor effects can be obscured. Existing analysis of covariance (ANCOVA) methods are typically based on the assumption of equal variances across the groups. These methods tend to not control the type-1 error rate satisfactorily under variance heteroscedasticity. A method numerically implemented in SAS has fitted two modified versions of ATS to ANCOVA models. However, clear theoretic framework has not been provided to support the SAS procedure. In our research, we tackle this issue and developed a new ATS based on the spirit of Behrens-Fisher problem and computed the degree of freedom by Box-type approximation in a general ANCOVA model. Extensive simulation studies show that our method is comparable to a SAS-based method, which controls the nominal type-1 error rate, even for very small sample sizes, moderately skewed distributions and under variance heteroscedasticity.

email: cong.cao1@utdallas.edu

SYSTEMS OF PARTIALLY LINEAR MODELS (SPLM) FOR MULTI-CENTER STUDIES

Lei Yang*, New York University
Yongzhao Shao, New York University

Multi-center studies are increasingly common in many medical and epidemiology investigations to borrow strength and increase sample size and statistical power. Frequently, it can be challenging to consistently assess the effect sizes of some important variables or biomarkers that are measured in different technical platforms in different centers. Systems of partially linear models (SPLM) are proposed here for modeling and analyzing data from multi-center study where the variables measured under different platforms in different centers are included in the non-linear terms of the SPLMs. Computing algorithms are introduced for simultaneous variable selection in SPLMs and characterization of homogeneous and heterogeneous effect of selected variables across different centers. Some theoretical properties are established including selection consistency and estimation consistency in the framework of reproducing kernel Hilbert space (RKHS) with L2 type and adaptive Lasso type penalty terms. The performance of the proposed method is further evaluated via simulation studies and using an Alzheimer's disease dataset.

email: ly888@nyu.edu

COMPARISON OF TWO TRANSFORMATION MODELS

Yuqi Tian*, Vanderbilt University
Bryan E. Shepherd, Vanderbilt University
Chun Li, Case Western Reserve University
Torsten Hothorn, University of Zurich
Frank E. Harrell, Vanderbilt University

Continuous response variables are often transformed to meet modeling assumptions, but the choice of transformation can be challenging. Two transformation models have been proposed: semiparametric cumulative probability models (CPM) and parametric most likely transformation models (MLT). Both approaches model the cumulative distribution function and require specifying a link function, which assumes the data follow a known distribution after a monotonic transformation. The two approaches estimate the transformation differently. With CPMs, an ordinal regression model is fit, treating each continuous response as its own ordered category and nonparametrically estimating the transformation. With MLTs, the transformation is parameterized using flexible basis functions. Conditional expectations and quantiles are derived. We compare them with simulations. Both have good performance with moderate to large sample sizes, but MLT slightly outperforms CPM in small sample sizes under correct models. CPM is more robust to model misspecifications, detection limits, and data rounding. Except in the simplest situations, both methods outperform basic transformation approaches commonly used.

email: yuqi.tian@vanderbilt.edu

ON EXTERNALLY CALIBRATING TIME-DEPENDENT ABSOLUTE RISK FOR TIME-TO-EVENT OUTCOME

Jiayin Zheng*, Fred Hutchinson Cancer Research Center
Li Hsu, Fred Hutchinson Cancer Research Center
Yingye Zheng, Fred Hutchinson Cancer Research Center

Accurate risk prediction for time-to-event outcome is valuable for population control of chronic diseases. It is of great interest to generalize risk prediction models to other populations. However, while hazard ratios of risk factors are often considered common across populations, the generalization of absolute risk can be problematic due to the potential differences between populations. We propose a novel statistical method to recalibrate the prediction model for the target population by incorporating the auxiliary information of the target population. A constrained likelihood is used to re-weight the cohort samples, allowing for the potential difference of the covariate distributions. Then by solving the re-weighted estimating equation, we gain efficiency when the covariate distributions are the same, while obtaining a robust estimator when the distributions differ. The weights essentially shift the sample distribution towards the target population and make the sample more representative of the target population. The asymptotic properties are established. Simulations studies and a real-data problem are used to illustrate the proposed method.

email: statzjy@gmail.com

A GENERAL INFORMATION CRITERION FOR MODEL SELECTION BASED ON EMPIRICAL LIKELIHOOD

Chixiang Chen*, The Pennsylvania State University
Rongling Wu, The Pennsylvania State University
Ming Wang, The Pennsylvania State University

Conventional likelihood-based information criteria for model selection rely on the specific distribution of data. However, the specification of the true distribution underlying increasingly heterogeneous data has proven to be challenging, thus limiting the model selection of these data. To address this issue, we propose a general information criterion framework. This framework is first derived from the asymptotic manner of the marginal distribution based on empirical likelihood, under which plug-in estimators derived from external estimating equations are used to calculate the information criterion. More importantly, such a framework is versatile by allowing model selection to be performed within various contexts, such as generalized linear model, generalized estimating equations, and penalized regression. Further, we establish the consistency property of the framework under mild conditions. Through extensive simulations studies, we also observe a better selection performance in selected cases, compared with main existing approaches. Finally, the utility and usefulness of the new framework are validated by a real example of the Atherosclerosis Risk in Communities Study.

email: chencxxy@psu.edu

ADJUSTING FOR PARTICIPATION BIAS IN CASE-CONTROL GENETIC ASSOCIATION STUDIES WITH GENOTYPE DATA SUPPLEMENTED FROM FAMILY MEMBERS: AN EMPIRICAL LIKELIHOOD-BASED ESTIMATING EQUATION APPROACH

Le Wang*, Villanova University
Zhengbang Li, Central China Normal University
Clarice Weinberg, National Institute of Environmental Health Sciences, National Institutes of Health
Jinbo Chen, University of Pennsylvania

Collection of genotype data in case-control genetic association studies may often be incomplete for reasons related to genes. Such non-ignorable missingness structure might result in biased association analyses, which has been a widespread concern in studies of many phenotypes. Chen et al. (2016) proposed to collect genetic information from family members and developed a maximum likelihood method for bias correction. In this study, we develop an estimating equation approach to analyzing data collected from this design that allows for adjustment of covariates. It jointly estimates odds ratio parameters for genetic association and missingness, where a logistic regression model is used to relate missingness with genotype and other covariates. We use genetic information of family members to infer the missing genotype data. In the estimating equation for genetic association parameters, we weight the empirical likelihood score function based on subjects with genotype data by the inverse probabilities that their genotypes are available. We studied large and finite sample performance of our method via simulation studies and applied it to a family-based study of breast cancer.

email: lwang0217@gmail.com

---

95. BAYESIAN APPROACHES TO HIGH DIMENSIONAL DATA

LATENT MIXTURES OF FUNCTIONS TO CHARACTERIZE THE COMPLEX EXPOSURE RELATIONSHIPS OF PESTICIDES ON CANCER INCIDENCE

Sung Duk Kim*, National Cancer Institute, National Institutes of Health
Paul S. Albert, National Cancer Institute, National Institutes of Health

Understanding the relationships between chemical exposure and cancer incidence is an important probable in environmental epidemiology. Individually, each pesticide might transmit small amounts of risk, but taken together, may pose substantial cancer risk. Importantly, these effects can be highly non-linear and can be in different directions. We develop an approach that models the simultaneous effect of all chemicals as the sum of nonlinear functions of each chemical. Since it is highly probable that many chemicals transmit small amounts of risk, and only taken together would we anticipate a sizable effect, we do not use traditional model selection approaches such as LASSO. Instead, we propose an approach in which individual effects are modeled as mixtures of non-linear functions to characterize the simultaneous effects of many different agents. We use state-of-the-art Bayesian methodology to estimate models with potentially large number of mixtures and use penalized likelihood approaches to choose an appropriate number of nonlinear functions. We illustrate this new methodology with data from cohort studies trying to address this important public health issues.

email: kims2@mail.nih.gov

PRIOR KNOWLEDGE GUIDED ULTRAHIGH-DIMENSIONAL VARIABLE SCREENING

Jie He*, University of Michigan
Jian Kang, University of Michigan

Variable screening is a powerful and efficient tool for dimension reduction under the ultrahigh dimensional setting. However, most of existing methods overlook useful prior knowledge (pre-selection, dependence, grouping and ranking) in specific applications. In this work, from Bayesian modeling perspectives we develop a unified variable screening procedure for regression models. We discuss different constructions of posterior mean screening (PMS) statistics to incorporate different types of prior knowledge according to the specific applications. With non-informative prior specifications, PMS is equivalent to high-dimensional ordinary least-square projections (HOLP). We establish the sure screening and sure consistency for PMS with different types of prior knowledge. We show that PMS is robust to prior misspecification; and when the prior knowledge provides correct information on summarizing the true parameter settings, PMS can substantially improve the selection accuracy compared to HOLP and other existing methods. We illustrate our methods on extensive simulation studies and an analysis of neruoimaging data.

email: jiehe@umich.edu

SEMIPARAMETRIC BAYESIAN KERNEL SURVIVAL MODEL FOR HIGHLY CORRELATED HIGH-DIMENSIONAL DATA

Lin Zhang*, Eli Lilly and Company (previously at Virginia Tech University)
Inyoung Kim, Virginia Tech University

Kernel machine models can model complex associations among elements (e.g. genes) within a set or among different sets (e.g. different gene pathways). Set-based analyses (e.g. pathway-based analyses) possess appealing advantages over element-based analyses (e.g. gene-based analyses), while distinguishing "important" elements from "less important" elements within "significant" sets also has wide applications. In this study, we propose an integrated two-stage sequential procedure to model both set effects (s-BKSurv) and element effects (g-BKSurv). We developed a two-stage sequential procedure to test both set effects and element effects. We also propose an integrated decision rule (i-gBF) that allows adjustment for multiple comparisons under a full Bayesian scenario to control family-wise error rate. We demonstrate the excellent testing capability of the proposed approach in terms of true positive rate, false positive rate, accuracy, and precision under various simulation settings and an application data set.

email: linzhang@vt.edu

BAYESIAN VARIABLE SELECTION IN HIGH-DIMENSIONAL EEG DATA USING SPATIAL STRUCTURED SPIKE AND SLAB PRIOR

Shariq Mohammed*, University of Michigan
Dipak Kumar Dey, University of Connecticut
Yuping Zhang, University of Connecticut

With the advent of modern technologies, it is increasingly common to deal with data of multi-dimensions in various scientific fields of study. In this paper, we develop a Bayesian approach for the analysis of high-dimensional neuroimaging data. We specifically deal with EEG data, where we have a matrix of covariates corresponding to each subject from either the alcoholic or control group. The matrix covariates have a natural spatial correlation based on the locations of the brain, and temporal correlation as the measurements are taken over time. We employ a divide and conquer strategy by building multiple local Bayesian models at each time point separately. We incorporate the spatial structure through the structured spike and slab prior, which has inherent variable selection properties. The temporal structure is incorporated within the prior by learning from the local model from the previous time point. We pool the information from the local models and use a weighted average to design a prediction method. We perform some simulation studies to show the efficiency of our approach and demonstrate the local Bayesian modeling with a case study on EEG data.

email: shariq.mohammed@uconn.edu

WEIGHTED DIRICHLET PROCESS MODELING FOR FUNCTIONAL CLUSTERING WITH APPLICATION IN MATCHED CASE-CROSSOVER STUDIES

Wenyu Gao*, Virginia Tech University
Inyoung Kim, Virginia Tech University

Matched case-crossover studies have become popular in epidemiological research. Comparisons of case and control statuses are made on the same subjects over different time periods. However, the response variables do not belong to exponential families, but have binary clustered features. The effects of matching covariates are unknown and individual subject may have its own characteristics which can create heterogeneous subpopulations among strata. Furthermore, there is seldom conjugacy in Bayesian paradigm, which provides difficulties in analyses, especially with a Dirichlet Process (DP) prior. DP prior has the automatic clustering property but the clustering results are usually unsatisfactory. In this talk, we would like to propose a weighted Dirichlet Process Mixture (WDPM) model to study the functional clustering behavior in matched case-crossover studies. WDPM model takes on more information from the predictors to calculate the weight from each candidate DP prior. Compared to the DP model, the extra information helps on the clustering results. The advantages of the WDPM prior are demonstrated through simulation studies and empirical data within epidemiology.

email: wenyu6@vt.edu

96. FUNCTIONAL DATA ANALYSIS METHODS

PROBABILISTIC K-MEAN WITH LOCAL ALIGNMENT FOR FUNCTIONAL MOTIF DISCOVERY

Marzia A. Cremona*, The Pennsylvania State University
Francesca Chiaromonte, The Pennsylvania State University

The aim is to address the problem of discovering functional motifs, i.e. typical "shapes" that may recur several times in a set of (multidimensional) curves, capturing important local characteristics of these curves. We formulate probabilistic K-mean with local alignment, a novel algorithm that leverages ideas from Functional Data Analysis (joint clustering and alignment of curves), Bioinformatics (local alignment through the extension of high similarity "seeds") and fuzzy clustering (curves belonging to more than one cluster, if they contain more than one typical "shape"). Our algorithm identifies shared curve portions, which represent candidate functional motifs in a set of curves under consideration. It can employ various dissimilarity measures in order to capture different shape characteristics. After demonstrating the performance of the algorithm on simulated data, we apply it to discover functional motifs in "Omics" signals related to mutagenesis and genome dynamics, exploring high-resolution profiles of different mutation rates in regions of the human genome where these rates are globally elevated.

email: mac78@psu.edu

A DECOMPOSABLE MODEL FOR ANALYZING MULTIVARIATE FUNCTIONAL DATA

Luo Xiao*, North Carolina State University

A decomposable functional principal component model for analyzing multivariate functional data is proposed. Compared to existing methods, the new model gives new insights how the different functions from the same subject are correlated and also identifies principal components that are unique to each function. A new scalar on function regression is also proposed using this decomposable model. The methods are illustrated with a simulation study and on multiple data sets including a Diffusion Tensor Imaging data.

email: lxiao5@ncsu.edu

MODEL TESTING FOR GENERALIZED SCALAR-ON-FUNCTION LINEAR MODELS

Stephanie T. Chen*, North Carolina State University
Luo Xiao, North Carolina State University
Ana-Maria Staicu, North Carolina State University

Scalar-on-function linear models are a common method for regressing a functional predictor on a scalar response. However, functional models are more difficult to estimate and interpret than traditional linear models, and may be unnecessarily complex for a data application. Hypothesis testing can be used to guide model selection by determining if functional predictors are necessary. We propose a generalized response-restricted likelihood ratio test for functional linear models with responses from exponential family distributions. We are motivated by a study that uses functional logistic regression to identify multiple sclerosis patients using diffusion tensor imaging, and demonstrate the efficacy of our method with a simulation study and data applications. Our method can be used for functional data to test predictors for no effect, necessity of a functional form, or any polynomial form.

email: stchen3@ncsu.edu

REGRESSION ANALYSES OF DISTRIBUTIONS USING QUANTILE FUNCTIONAL REGRESSION

Hojin Yang*, University of Texas MD Anderson Cancer Center
Veerabhadran Baladandayuthapani, University of Michigan
Jeffrey S. Morris, University of Texas MD Anderson Cancer Center

We present methods to model the entire marginal distribution via the quantile function as functional data, regressed on a set of predictors. We call this approach quantile functional regression, regressing subject-specific marginal distributions across repeated measurements on a set of covariates, allowing us to assess which covariates are associated with the distribution in a global sense, as well as to identify distributional features characterizing these differences, including mean, variance, and various quantiles. To account for smoothness in the quantile functions, we introduce custom basis functions we call quantlets that are sparse, regularized, near-lossless, and empirically defined, adapting to the features of a given data set. We fit this model using a Bayesian framework that uses nonlinear shrinkage of quantlet coefficients to regularize the functional regression coefficients and provides fully Bayesian inference after fitting a Markov chain Monte Carlo. We demonstrate the benefit of the basis space modeling through simulation studies, and apply the method to Magnetic resonance imaging based radiomic dataset from Glioblastoma Multiforme.

email: hojiny0504@gmail.com

A BAYESIAN MODEL FOR CLASSIFICATION AND SELECTION OF FUNCTIONAL PREDICTORS USING LONGITUDINAL MRI DATA FROM ADNI

Asish K. Banik*, Michigan State University
Taps Maiti, Michigan State University

The objective of this paper is to use significant number of longitudinal volumetric MRI data as statistical predictors while predicting the probability of a certain patient belongs to Alzheimer's group or not. We dealt with logistic regression in Bayesian setup with large number of functional predictors. The direct sampling of regression coefficients from Bayesian logistic model is difficult due to its complicated likelihood function. Our aim was to avoid the complicated Metropolis-Hastings and develop an easily implementable Gibbs Sampler. The Bayesian estimation provides proper estimates of the model parameters which are also useful for building inference. Logistic regression calculates the log of odds of relative risk of AD compared to CN based on the selected longitudinal predictors rather than just classifying patients. We use 115 functional predictors which are nothing but volume measurements of different sub-regions of whole brain. Spike-and-Slab prior ensures that a large number redundant predictors are dropped from the model. The sub-regions which are selected by our method are very important for future studies to detect the progression of dementia.

email: banikasi@stt.msu.edu

MULTILEVEL LOCALIZED-VARIATE PCA FOR CLUSTERED MULTIVARIATE FUNCTIONAL DATA

Jun Zhang*, University of Pittsburgh
Greg J. Siegle, University of Pittsburgh
Robert T. Krafty, University of Pittsburgh

In this talk, we discuss multilevel localized-variate functional principal component analysis (MLVFPCA) for extracting interpretable basis functions that account for intra and inter-cluster variability in a clustered multivariate process. The methodology is motivated by popular neuroscience experiments where patients' brain activity is recorded using EEG or fMRI, summarized as power within multiple time-varying frequency bands, in multiple brain regions, while completing multiple tasks. The basis functions found by MLVFPCA can be both localized within a variate (i.e. nonzero only within a subinterval of a frequency band) and sparse among variates (i.e. zero across an entire frequency band). The sparsity is achieved by rank-one based convex optimization with matrix L1 and block Frobenius norm based penalties. To jointly model data across electrodes and tasks, we decompose the functional variability into subject level, task level and electrode level variability. The application of MLVPFCA to a study of emotional response in patients diagnosed with affect disorder reveals new insights into blunted affect, or mechanisms of "emotionally shutting down".

email: juz30@pitt.edu

97. NEXT GENERATION SEQUENCING

A PROBABILISTIC MODEL TO ESTIMATE THE TEMPORAL ORDER OF PATHWAY MUTATIONS DURING TUMORIGENESIS

Menghan Wang*, University of Kentucky
Chunming Liu, University of Kentucky
Chi Wang, University of Kentucky
Arnold Stromberg, University of Kentucky

Cancer arises through accumulation of somatically acquired genetic mutations. An important question is to understand the temporal order of mutations during tumorigenesis, since early mutations may indicate potential therapeutic targets and late mutations may be involved in metastasis.In this paper, we develop a novel statistical method to estimate the temporal order of mutations in biological

pathwayswhile accounting for the difference in mutations' functional impacts. A probabilistic model is constructed for each pair of biological pathways to characterize the probability of mutational events from those two pathways occurring in a certain order. The functional impact of each mutation, quantified by PolyPhen-2 score, is incorporated into the model to weigh more on the mutation that is more likely to be damaging. A maximum likelihood method is used to estimate model parameters and infer the probability of one pathway being mutated prior to the other. Simulation studies and analysis of mutation data from The Cancer Genome Atlas demonstrate that our method is able to accurately estimate the temporal order of pathway mutations.

email: mwa287@g.uky.edu

PROBABILISTIC INDEX MODELS FOR TESTING DIFFERENTIAL GENE EXPRESSION IN SINGLE-CELL RNA SEQUENCING (scRNA-Seq) DATA

Alemu Takele Assefa*, Ghent University, Belgium
Olivier Thas, Ghent University, Belgium
Jo Vandesompele, Ghent University, Belgium

Analysis of scRNA-seq data is of paramount interest in studying diseases such as cancer. For example, it has been used to characterize intra-tumoral heterogeneity and to classify tumor sub-populations. It is often used for testing for differential expression (DE), genes with different expression across conditions of interest. However, the particular characteristics of scRNA-seq data, including overdispersion and dropout events, are challenging issues for the performance of statistical tools. Studies have demonstrated that classical non-parametric tests can be used to robustly detect DE. However, these tests lose power due to the library size variability and the large frequency of zero counts. Moreover, they cannot be used for complex experimental designs. We propose a semi-parametric approach based on Probabilistic Index Models, which form a class of models that generalizes classical rank tests. Our method does not rely on strong distributional assumptions and it accounts for library size variability. Our simulation studies demonstrate that our tests for DE succeeds well in controlling the FDR, while showing good sensitivity as compared to competitor methods.

email: AlemuTakele.Assefa@UGent.BE

DETECTION OF DIFFERENTIALLY EXPRESSED GENES IN DISCRETE SINGLE-CELL RNA SEQUENCING DATA USING A HURDLE MODEL WITH CORRELATED RANDOM EFFECTS

Michael Sekula*, University of Louisville
Jeremy Gaskins, University of Louisville
Susmita Datta, University of Florida

Single-cell RNA sequencing (scRNA-seq) technologies are revolutionary tools allowing researchers to examine gene expression at the level of a single cell. Traditionally, transcriptomic data have been analyzed from bulk samples, masking the heterogeneity now seen across individual cells. Even within the same cellular population, genes can be highly expressed in some cells but not expressed (or lowly expressed) in others. Therefore, the computational approaches used to analyze bulk RNA sequencing data are not appropriate for the analysis of scRNA-seq data. Here, we present a novel statistical model for high dimensional and zero-inflated scRNA-seq count data to identify differentially expressed genes across cell types. Correlated random effects are employed based on an initial clustering of cells to capture the cell-to-cell variability within treatment groups. Moreover, this model is flexible and can be easily adapted to an independent random effect structure if needed. Both simulated and real data are utilized to demonstrate how our model outperforms other popular methods designed for detecting differentially expressed genes.

email: michael.sekula@louisville.edu

TRANSFER LEARNING FOR CLUSTERING ANALYSIS FROM SINGLE-CELL RNA-Seq DATA

Jian Hu*, University of Pennsylvania, Perelman School of Medicine
Xaingjie Li, Renmin University of China
Gang Hu, Nankai University
Mingyao Liu, University of Pennsylvania, Perelman School of Medicine

Recent development of single-cell RNA-seq technologies has led to biological discoveries yet also introduced statistical and computational challenges. A vital step in single-cell RNA-seq analysis is cell type clustering. Existing methods suffer from low accuracy when dataset has few cells or low sequencing depth. To overcome this limitation,we want to utilize knowledge from a large training dataset to help cluster on a relatively small targeting dataset. As many single-cell studies are multi-year projects, it is appealing to transfer cell type knowledge learned from old data to a new dataset generated in future years. Therefore, transfer learning suits perfectly for continuously generated data in these single-cell studies. We explored a transfer learning method to do cell clustering using single-cell RNA-seq data. In this method, one network developed for a task using training data is reused as the initial point for the clustering network on a second task using targeting data. To evaluate our method, we analyzed multiple human pancreas single-cell RNA-seq datasets. The clustering accuracy and efficiency were greatly improved if we utilize transferred information from training dataset.

email: jianhu@pennmedicine.upenn.edu

INCORPORATING SINGLE-CELL RNA-Seq DATA TO INFER ALLELE-SPECIFIC EXPRESSION

Jiaxin Fan*, University of Pennsylvania, Perelman School of Medicine
Rui Xiao, University of Pennsylvania, Perelman School of Medicine
Mingyao Li, University of Pennsylvania, Perelman School of Medicine

Allele-specific expression (ASE) can be quantified by the relative expression of two alleles in a diploid individual, and such expression imbalance may explain phenotypic variation and disease pathophysiology. Existing methods detect ASE using easily obtainable bulk RNA-seq data, a data type that averages out possible heterogeneity in a mixture of different cell types. Since ASE may vary across different cell types, with the recent advance in single-cell RNA sequencing (scRNA-seq), characterizing ASE at the cell type resolution may help reveal more about the gene regulation. However, scRNA-seq data is costly to generate and noisy with excessive zeros due to transcriptional bursting. Therefore, it is desirable to incorporate information obtained from scRNA-seq data together with bulk data to infer cell type specific ASE. By employing cell type deconvolution and simultaneously modelingof multi-individual

information, we are able to detect cell type specific ASE. Extensive simulations indicate that our method performsconsistently well under a variety of scenarios.

email: jiaxinf@pennmedicine.upenn.edu

A MINIMAX OPTIMAL TEST FOR RARE-VARIANT ANALYSIS IN WHOLE-GENOME SEQUENCING STUDIES

Yaowu Liu*, Harvard University
Xihong Lin, Harvard University

Optimality criteria in hypothesis testing, such as unbiasness, invariance and asymptotic powerfulness, do not take the signal strength into consideration, which, however, can have a great impact on the power of tests. As signal strength is ignored in the construction, many existing tests, including the classic F test, Hotelling's T2 test, and the empirical Bayes based score (EB-score) test, can have substantial loss of power in the presence of weak/moderate signal strength. We defined a novel risk function that takes the signal strength into account and proposed a new test that is minimax optimal with respect to the defined risk function within a class of tests that include the F test and the EB-score test. The new test does not sacrifice the simplicity compared to classical tests, i.e., the test also has a simple form and its p-value can be calculated analytically and efficiently. We applied the new test to the analysis of whole genome sequencing data from the Atherosclerosis Risk in Communities (ARIC) study and it identified 40%--200% more significant regions than the F test and the Sequence Kernel Association Test (SKAT) that is a EB-score test.

email: yaowuliu@hsph.harvard.edu

BAMM-SC: A BAYESIAN MIXTURE MODEL FOR CLUSTERING DROPLET-BASED SINGLE CELL TRANSCRIPTOMIC DATA FROM POPULATION STUDIES

Zhe Sun*, University of Pittsburgh
Ying Ding, University of Pittsburgh
Wei Chen, Children's Hospital of Pittsburgh of UPMC, University of Pittsburgh
Ming Hu, Cleveland Clinic Foundation

The recently developed droplet-based single cell transcriptome sequencing (scRNA-seq) technology makes it feasible to perform a population-scale scRNA-seq study, in which the transcriptome is measured for tens of thousands of single cells from multiple individuals. Despite the advances of many clustering methods, there are few tailored methods for population-scale scRNA-seq studies. Here, we have developed a BAyesian Mixture Model for Single Cell sequencing (BAMM-SC) method to cluster scRNA-seq data from multiple individuals simultaneously. Specifically, BAMM-SC takes raw data as input and can account for data heterogeneity and batch effect among multiple individuals in a unified Bayesian hierarchical model framework. Results from extensive simulations and application of BAMM-SC to in-house scRNA-seq datasets using blood, lung and skin cells from humans or mice demonstrated that BAMM-SC outperformed existing clustering methods with improved clustering accuracy and reduced impact from batch effects. BAMM-SC has been implemented in a user-friendly R package.

email: zhs31@pitt.edu

---

98. COMPETING RISKS AND CURE MODELS

GENERAL REGRESSION MODEL FOR THE SUBDISTRIBUTION OF A COMPETING RISK UNDER LEFT-TRUNCATION AND RIGHT-CENSORING

Anna Bellach*, University of Washington
Michael R. Kosorok, University of North Carolina, Chapel Hill
Peter Gilbert, Fred Hutchinson Cancer Research Center
Jason P. Fine, University of North Carolina, Chapel Hill

Left-truncation poses additional challenges for the analysis of complex time to event data. We propose a general semiparametric regression model for left-truncated and right-censored competing risks data. Targeting the subdistribution hazard, our parameter estimates are directly interpretable with regard to the cumulative incidence function. Our approach accommodates external time dependent covariate effects on the subdistribution hazard. We establish consistency and asymptotic normality of the estimators and propose a sandwich estimator of the variance. In comprehensive simulation studies we demonstrate a solid performance of the proposed method, thereby comparing the sandwich estimator to the inverse Fisher information. Applying the new method to HIV-1 vaccine efficacy trial data we investigate how participant factors associate with the time from adulthood until HIV-1 infection.

email: abellach@uw.edu

DOUBLY ROBUST OUTCOME WEIGHTED LEARNING ESTIMATOR FOR COMPETING RISK DATA WITH GROUP VARIABLE SELECTION

Yizeng He*, Medical College of Wisconsin
Mi-Ok Kim, University of California, San Francisco
Soyoung Kim, Medical College of Wisconsin
Kwang Woo Ahn, Medical College of Wisconsin

Competing risks data are commonly encountered in medical investigations. It occurs when each study subject can experience one and only one of several distinct types of events or failures. Often, finding the optimal treatment regime that gives an individual patient the best outcome under the competing risks settings is of interest, but the traditional regression methods are lack of robustness. In this paper, we propose a doubly robust estimator, driven by the idea from outcome weighted learning, for the expectation of a function of failure time from the cause of interest, given treatment and covariates. We show, through simulation studies, the proposed estimator is consistent if either the propensity score for the treatment and censoring distribution are correctly specified, or the distribution of failure time from cause 1 given treatment and covariates and censoring distribution are correctly specified. Group variable selection is also performed using the support vector machine.

email: yizhe@mcw.edu

COVARIATE ADJUSTMENT FOR TREATMENT EFFECT ON COMPETING RISKS DATA IN RANDOMIZED CLINICAL TRIALS

Youngjoo Cho*, University of Wisconsin, Milwaukee
Cheng Zheng, University of Wisconsin, Milwaukee
Mei-Jie Zhang, Medical College of Wisconsin

The double blinded randomization trial is a gold standard for estimating average causal effect (ACE). It does not require adjustment for covariates. However, in most case, adjustment of covariates that are strong predictor of the outcome could improve efficiency for the estimation of ACE. But when covariates are high-dimension, adjust all covariates in the model will lose efficiency or worse, lose identifiability. Recent work has shown that for linear regression, an estimator under risk consistency (e.g., LASSO, Random Forest) for the regression coefficients could always lead to improvement in efficiency. In this work, we studied the behavior of adjustment estimator for competing risk data analysis. Simulation study shows that the covariate adjustment provides the more efficient estimator than unadjusted one.

email: yvc5154@gmail.com

MARGINAL CURE RATE MODELS FOR LONG-TERM SURVIVORS

Jianfeng Chen*, Kansas State University
Wei-Wen Hsu, Kansas State University
David Todem, Michigan State University
KyungMann Kim, University of Wisconsin, Madison

Two-component mixture models for long-term survivors, known as standard cure rate models, have been widely used and intensively discussed in the literature. Much of attention has been put on to understand the covariate effects on both the cure fraction and hazard rate components of this model. However, it is extremely challenging to interpret the covariate effects on the overall survival response when these covariates are shared by the cure fraction and hazard rate simultaneously. In this paper, we propose a marginal cure rate model that can provide a general framework to investigate the covariate effects on the survival outcome of the whole population. Technically, a novel transformation that can relate the covariates directly to the marginal mean hazard rate is adopted and embedded in the likelihood function of a standard cure rate model. The proposed marginal model is evaluated through extensive simulation studies and illustrated with an application to the liver cancer data from the SEER registry.

email: sin16terry@ksu.edu

AN APPLICATION OF THE CURE MODEL TO A CARDIOVASCULAR CLINICAL TRIAL

Varadan V. Sevilimedu*, Department of Veteran Affairs and Yale University
Shuangge Ma, Yale University
Pamela Hartigan, Department of Veteran Affairs
Tassos C. Kyriakides, Department of Veteran Affairs

Intermediate events play an important role in determining the risk of developing a medical condition over time and should thus be accounted for in the context of survival analysis. Myocardial infarction (MI) is one such medical condition whose hazards are also dependent upon the occurrence of an intermediate event such as acute coronary syndrome (ACS). The study of the role that ACS plays in altering the hazards of MI becomes complicated when there is a cure fraction in the population. Data from the Clinical Outcomes Utilizing Revascularization and Aggressive Drug Evaluation (COURAGE) trial provide the scenario where the existence of a cure fraction is highly likely. In this study we assess the role of ACS in altering the pathway towards developing an MI, in the presence of a cure fraction. We adapt non-parametric maximum likelihood estimation to estimate the coefficients of this multi-part cure model. Simulation studies show that the estimates have good asymptotic properties. In addition, we also utilize this dataset to explore the use of a proportionality constraint to help reduce the dimensionality of this multi-part model. Future direction in analysis is also presented.

email: varadan.sevilimedu@gmail.com

ESTIMATION THE SERIAL INTERVAL USING CURE MODELS

Laura F. White*, Boston University
Helen E. Jenkins, Boston University
Paola Sebastiani, Boston University
Yicheng Ma, Boston University

The serial interval is the time between onset of symptoms in an infector/infectee and is critical for modeling studies, control measures, and understanding transmission dynamics. Methods to estimate this distribution frequently fail to account for censoring and erroneously assume that exposed and censored individuals all develop disease. We show how cure models can be used to account for censoring and cure among exposed individuals. We describe a Bayesian approach for implementing the cure model, by placing a prior on the cure rate using known literature. A simulation study shows when this method provides advantages over more naive methods and we provide diagnostic tools for model selection. We apply this to Brazilian serial interval data. In simulation, bias reduces ten-fold as the cure rate decreases from 97% to 20%. Cure models with a flat prior decrease bias by between 75% and 99% as the cure rate decreases, relative to a naive MLE approach. In Brazil, we estimate a mean serial interval of 2.4 years (95% CI: 1.7-3.4 years) for the naive model and 5.3 years (95% CI: 2.3-20.7 years) with a cure model. Cure models are less biased with even moderate cure rates and censoring.

email: lfwhite@bu.edu

99. MONITORING HEALTH BEHAVIORS WITH MULTI-SENSOR MOBILE TECHNOLOGY

VARIABLE SELECTION IN THE CONCURRENT FUNCTIONAL LINEAR MODEL

Jeff Goldsmith*, Columbia University
Joseph E. Schwartz, Columbia University Medical Center and Stony Brook University

We develop methods for variable selection when modeling the association between a functional response and functional predictors that are observed on the same domain. This data structure, and the need for such methods, is exemplified by our motivating example: a study in which blood pressure values are observed throughout the day together with measurements of physical activity, heart rate, location, posture, attitude, and other quantities that may influence blood pressure. We estimate the

coefficients of the concurrent functional linear model using variational Bayes and jointly model residual correlation using functional principal components analysis. Latent binary indicators partition coefficient functions into included and excluded sets, incorporating variable selection into the estimation framework. The proposed methods are evaluated in simulated- and real-data analyses.

email: ajg2202@cumc.columbia.edu

STATISTICAL MODELLING OF CROSS-SYSTEMS BIOMARKERS

Vadim Zipunnikov*, Johns Hopkins Bloomberg School of Public Health
Haochang Shou, University of Pennsylvania
Mike Xiao, National Institute of Mental Health, National Institutes of Health
Kathleen Merikangas, National Institute of Mental Health, National Institutes of Health

Daily electronic diaries available through smart-phones/watches are now extensively used for ecological momentary sampling that taps patterns of many homeostatic systems including sleep, emotional states, dietary intake and others to assess behavioral components of human homeostatic systems. We develop a statistical framework that focuses on joint modelling the complex interaction of multiple biomarkers of sleep, physical activity, and circadian system coupled with the context from electronic diaries to characterize multi-modal behavioral phenotypes. We apply the approach to the data from Family Study of Mood Disorders that collected EMA and actigraphy data on 315 subjects. We will illustrate how the approach provides a deeper understanding of the dynamic interplay of multiple brain-body systems and will facilitate the development of personalized in-time prevention strategies.

email: vzipunni@jhsph.edu

TRANSLATIONAL BIOMARKERS FOR QUALITY OF SLEEP

Dmitri Volfson*, Takeda Pharmaceutical Company
Brian Tracey, Tufts University
Tamas Kiss, Hungarian Academy of Sciences
Derek Buhl, Takeda Pharmaceutical Company

Here we present two collaborative efforts utilizing supervised learning to develop tools for preclinical and early clinical biomarker discovery, which provide robust quantification of translatable biomarkers derived from the sleep architecture. First, we describe a robust automated method for polysomnography utilizing EEG & EMG signals. It works across species, mice, rats and non-human primates, and across experiments and labs. Second, we present an approach which combines measurements of activity and heart rate variability to improve sleep-wake activity endpoints and enable scalable longitudinal monitoring in clinical trials.

email: dnvolfson@gmail.com

STATISTICAL MODELING FOR INTEGRATING DATA FROM MULTIPLE WEARABLE SENSORS TO DETECT AFFECT LABILITY

Fengqing Zhang*, Drexel University
Tinashe Tapera, Drexel University
Adrienne Juarascio, Drexel University

Emotion dysregulation is a broad, transdiagnostic risk and maintenance factor for numerous psychological disorders. Current treatments targeting emotional regulation are only partly effective, likely due to suboptimal delivery systems. In this study, we take the smartphone-based just-in-time, adaptive intervention approach, allowing real-time data collection and in-the-moment prediction of mood or behavioral states of interest. Various sensor data such as heart rate, physical activity, skin temperature and electrodermal activity are collected. The combination of multiple wearable sensor data provides a more comprehensive picture of the research problem. However, it also requires the development of new statistical methods that can make full usage of the informational complexity. A large number of physiological features are extracted from different wearable sensors in both time and frequency domains. We then develop a new data integration strategy based on multiple kernel learning to combine features from different sensors to predict affect lability. Analysis results show that the proposed approach can effectively and efficiently detect urges and emotional eating episodes.

email: fengqingzoezhang@gmail.com

100. CURRENT METHODS TO ADDRESS DATA ERRORS IN ELECTRONIC HEALTH RECORDS

MULTIPLE IMPUTATION TO ADDRESS DATA ERRORS IN ELECTRONIC HEALTH RECORD ANALYSES: ADVANTAGES AND DISADVANTAGES

Bryan E. Shepherd*, Vanderbilt University
Mark J. Giganti, Harvard University

Routinely collected electronic health record (EHR) data are increasingly used for medical research. However, the quality of EHR data is often poor, and results that use them can be misleading. One approach to address EHR data errors is to validate data in a subsample of records, model relationships between validated and unvalidated data, and then from these models to multiply impute validated data in the vast majority of records that were not validated. This multiple imputation approach has several appealing features including its simplicity, its capacity to handle complicated multivariable errors, its ability to address errors in inclusion/exclusion criteria, and its facility to be incorporated into a wide range of statistical methods. However, there are challenges to using multiple imputation in this context, including bias due to model mis-specification and incompatible imputation and analysis models that generally lead to overestimation of the variance of estimators. Advantages and disadvantages of using multiple imputation with EHR data, as well as strategies for overcoming some of these challenges, will be illustrated using data from an HIV cohort.

email: bryan.shepherd@vanderbilt.edu

RAKING AND REGRESSION CALIBRATION: METHODS TO ADDRESS BIAS INDUCED FROM CORRELATED COVARIATE AND TIME-TO-EVENT ERROR

Eric J. Oh*, University of Pennsylvania
Pamela A. Shaw, University of Pennsylvania

Medical studies that depend on electronic health records data are often subject to measurement error, as the data were not collected to support research questions under study. There are many methods to address covariate measurement error; however, error in time-to-event outcomes has also been shown to cause significant bias and methods to address it are relatively underdeveloped. More generally, it is common to observe errors in both the covariate and outcome that are correlated. We propose regression calibration estimators to address both time-to-event and correlated covariate and time-to-event error. It is well known that regression calibration estimators in time-to-event settings are biased. Thus, we additionally propose raking estimators for the two-phase sampling design, where auxiliary variables are collected at the first phase to improve efficiency for estimation based on the precise data available on a subset in the second phase. Simulation studies are presented to examine the performance of the proposed estimators under varying parameter settings. The methodology is illustrated on an observational HIV cohort from the Vanderbilt Comprehensive Care Clinic.

email: ericoh@pennmedicine.upenn.edu

AN AUGMENTED ESTIMATION PROCEDURE FOR EHR-BASED ASSOCIATION STUDIES ACCOUNTING FOR DIFFERENTIAL MISCLASSIFICATION

Yong Chen*, University of Pennsylvania
Jiayi Tong, University of Pennsylvania
Jing Huang, University of Pennsylvania
Xuan Wang, Zhejiang University
Jessica Chubak, Kaiser Permanente Washington Health Research Institute
Rebecca Hubbard, University of Pennsylvania

Electronic health records (EHR) have been widely used in various healthcare and medical related areas to investigate research questions. Generally, phenotypes of patients are derived from a high-throughput phenotyping algorithm, whereas chart reviews (deemed as a gold standard for the true phenotype) are available only for a small subset of patients. In this talk, I will present a simple estimation procedure to improve the statistical efficiency of estimated effect sizes in EHR-based association studies. The key idea is to leverage the error-prone EHR-derived phenotypes to augment the statistical efficiency of the estimated associations. An appealing property of our method is that it can consistently improve the statistical efficiency in both differential and non-differential misclassification settings, without modeling the dependency of misclassification rates on exposure level. Our method is a data-driven approach in the sense that the complexity of differential/non-differential misclassification is handled non-parametrically. This is a joint work with Jiayi Tong, Jing Huang, Jessica Chubak, Xuan Wang and Rebecca Hubbard.

email: ychen123@pennmedicine.upenn.edu

101. FINDING THE RIGHT ACADEMIC FIT: EXPERIENCES FROM FACULTY ACROSS THE ACADEMIC SPECTRUM

PANEL DISCUSSANTS:

Jianwen Cai, University of North Carolina, Chapel Hill
Alexandra Hanlon, University of Pennsylvania
Leslie McClure, Dornsife School of Public Health at Drexel University
Sujata Patil, Memorial Sloan Kettering Cancer Center
Randall H. Rieger, West Chester University

102. NOVEL INTEGRATIVE OMICS APPROACHES FOR UNDERSTANDING COMPLEX HUMAN DISEASES

INTEGRATIVE ANALYSIS OF INCOMPLETE MULTI-OMICS DATA

Danyu Lin*, University of North Carolina, Chapel Hill

Recent technological advances have made it possible to collect multiple types of omics data on a large number of individuals. Although genotypes are typically available for all study subjects, other data types may be measured only on a subset of subjects due to cost constraints. In addition, quantitative omics measurements are subject to detection limits. We propose a simple and powerful approach to jointly analyze such incomplete multi-omics data. We relate the quantitative omics variable to genetic variants through a linear regression model and relate the phenotype to the quantitative omics variable through a generalized linear model. We derive the joint likelihood for the two models by allowing the quantitative omics variable to be potentially missing and subject to detection limits. We carry out maximum likelihood estimation through an efficient EM algorithm and provide appropriate variance estimation. We demonstrate the advantages of the proposed methods over imputation methods through extensive simulation studies and provide an application to a chronic obstructive lung disease study measuring 100 blood proteins and half a million SNPs on 3,000 patients.

email: lin@bios.unc.edu

AN INTEGRATIVE FRAMEWORK TO EMPOWER GENOMICS-INFORMED ANALYSIS OF WHOLE GENOME SEQUENCING DATA FOR COMPLEX DISEASES

Bingshan Li*, Vanderbilt University

For most of complex diseases the majority of the genetic variants remain undiscovered after a decade long genome-wide association studies (GWAS). Whole genome sequencing (WGS) is becoming standard for genetics studies of complex traits, and large-scale studies such as the Trans-Omics for Precision Medicine (TOPMed) and the Genome Sequencing Program (GSP), are currently sequencing >200,000 whole genomes of several complex traits and diseases. A key challenge is that most risk variants are in noncoding genome, and the genetics mechanisms remain elusive given that their target genes are generally unknown. In this study we developed a Bayesian framework to select risk genes from known GWAS loci, and then used machine-learning approaches to prioritize novel risk genes across the genome. We further developed a statistical method to obtain gene-level statistical significance with increased power by taking into account the ranking of the genes

across the genome. We applied this integrative framework to WGS and GWAS data of autism and schizophrenia to identify novel genes implicated in such psychiatric disorders in a genomics-informed manner.

email: bingshan.li@vanderbilt.edu

PROBABILISTIC TWO SAMPLE MENDELIAN RANDOMIZATION FOR GENOME-WIDE ASSOCIATION STUDIES

Xiang Zhou*, University of Michigan
Zhongshang Yuan, University of Michigan

Two sample Mendelian randomization analyses have been widely applied in various genome-wide association studies to infer the causal relationship between omics phenotypes and complex traits. A key modeling assumption of these Mendelian randomization analyses is that the instrumental variables do not have horizontal pleiotropy effects -- an assumption that is challenging to validate or control for in real data applications. Here, we propose a probabilistic version of the commonly used Egger regression to test and control for horizontal pleiotropic effects in Mendelian randomization studies. A key feature of our method is its ability to accomodate for high-dimensional instrumental variables. With extensive simulations, we show that our method provides calibrated type I error control and is more powerful than several existing approaches in detecting causal associations between omics phenotypes and complex traits. Finally, we illustrate the benefits of our method in applications to three large-scale genome-wide association studies including the UK Biobank.

email: xzhousph@umich.edu

A SEMI-SUPERVISED APPROACH FOR PREDICTING CELL-TYPE SPECIFIC FUNCTIONAL CONSEQUENCES OF NON-CODING VARIATION USING MPRAs

Zihuai He*, Stanford University
Linxi Liu, Columbia University
Kai Wang, Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia
Iuliana Ionita-Laza, Columbia University

Predicting the functional consequences of genetic variants in non-coding regions is a challenging problem. We propose here a semi-supervised approach, GenoNet, to jointly utilize experimentally confirmed regulatory variants (labeled variants), millions of unlabeled variants genome-wide, and more than a thousand cell/tissue type specific epigenetic annotations to predict functional consequences of non-coding variants. Through the application to several experimental datasets, we demonstrate that the proposed method significantly improves prediction accuracy compared to existing functional prediction methods at the tissue/cell type level, but especially so at the organism level. Importantly, we illustrate how the GenoNet scores can help in fine-mapping at GWAS loci, and in the discovery of disease associated genes in sequencing studies. As more comprehensive lists of experimentally validated variants become available over the next few years, semi-supervised methods like GenoNet can be used to provide increasingly accurate functional predictions for variants genome-wide and across a variety of cell/tissue types.

email: statzihuai@gmail.com

103. TEACHING DATA SCIENCE THROUGH CASE-STUDIES

MOTIVATING DATA SCIENCE THROUGH CASE STUDIES IN PUBLIC HEALTH

Leah R. Jager*, Johns Hopkins Bloomberg School of Public Health

An increased demand for training in statistics, biostatistics, and data science among students with diverse backgrounds and degree interests gives faculty an opportunity to meet students where they are in terms of both ability and contextual interests. At Johns Hopkins, we teach introductory biostatistics in a public health context to an audience of undergraduate Public Health majors. The course is divided into modules, each motivated by a public health question. At the end of each module, students are guided through an analysis to address the question and then summarize their findings in a written report. In this way, students work through an in-depth case study in a context of interest to them. This case-study framework teaches the student to make important connections between the scientific question, the data, and the relevant statistical concepts through hands-on data analysis. However, constructing meaningful case-studies are time intensive. I'll share experiences teaching this course and information about an Open Case Studies resource that we are developing to allow educators to easily incorporate case studies into their own courses.

email: ljager@jhu.edu

TEACHING GENOMIC DATA SCIENCE: SUMMARIZATION, EXPLORATION, AND REPRODUCIBILITY

Michael I. Love*, University of North Carolina, Chapel Hill

Biologists and biomedical researchers working in genomics are now regularly generating high-dimensional, multi-assay datasets which they themselves may not have the necessary training to analyze. Trainees in Biology or Genetics, or in biomedical research institutes, may need to seek data science courses offered by Biostatistics or Statisics Departments in order to gain the skills necessary to extract useful information and make correct inference from their genomic datasets. I will discuss strategies in teaching genomic data science to such investigators and trainees, including (1) conceptual frameworks for processing of genomic data into manageable and relevant summaries, including uncertainty measures, (2) encouraging exploratory data analysis of high dimensional genomic data for quality control and discovery, and (3) teaching the benefits and tools necessary for reproducible workflows and code repositories as companions for publications involving analysis of genomic datasets.

email: michaelisaiahlove@gmail.com

BEFORE TEACHING DATA SCIENCE, LET'S FIRST UNDERSTAND HOW PEOPLE DO IT

Rebecca Nugent*, Carnegie Mellon Statistics & Data Science
Philipp Burckhardt, Carnegie Mellon Statistics & Data Science
Ronald Yurko, Carnegie Mellon Statistics & Data Science

The first wave of Data Science programs was largely built on a foundation of already existing computing-oriented classes; less effort was spent on how diverse backgrounds and disciplines approach data science. At Carnegie Mellon, the

Department of Statistics & Data Science teaches thousands of students with degrees ranging from Pre-Med to Rhetoric to Chemistry to Business to Statistics & Machine Learning and is well positioned to tackle this pedagogical challenge. We present ISLE (Interactive Statistics Learning Environment), an interactive platform that removes the computing cognitive load and lets students explore Statistics & Data Science concepts in both structured and unstructured ways. The platform also supports student-driven inquiry and case studies. We track and model every click, word used, and decision made along the data analysis pipeline from loading the data to the final written report. The platform is flexible enough to allow adaptation, providing different modes of data analysis instruction, active learning opportunities, and exercises for different subsets of the population. Teaching Data Science while simultaneously learning how we do it.

email: rnugent@andrew.cmu.edu

INTRODUCTION TO DATA SCIENCE, CASE-BY-CASE

Mine Cetinkaya-Rundel*, Duke University and RStudio

Interest in data science is surging, which means nowadays it's pretty easy to fill the seats in an introductory data science class. But how do we effectively take the students through a challenging curriculum once they are in the class? We argue that the answer is an application first approach where the curriculum is divided into learning modules, each covering a batch of connected learning goals and designed around a case study. In this talk we present the curriculum for such a course intended for an audience of Duke University students with little to no computing or statistical background, and focuses on data wrangling, exploratory data analysis, data visualization, and effective communication. This course serves not as a first and thorough exposure to computing essentials for data science (including programming with R, reproducibility with R Markdown, and version control and collaboration with git/GitHub) but also as a gateway for the statistical science major. We will discuss in detail the course design philosophy and pedagogical considerations as well as give examples from the case studies used in the course.

email: mine@stat.duke.edu

104. NONCONVEX OPTIMIZATION AND BIOLOGICAL APPLICATIONS

LOCAL FALSE DISCOVERY RATES FOR NONCONVEX PENALTIES IN HIGH-DIMENSIONAL REGRESSION MODELS

Patrick Breheny*, University of Iowa
Ryan E. Miller, University of Iowa

Nonconvex penalties such as MCP and SCAD have been shown to have a number of advantages for high-dimensional regression modeling, both theoretically and in practice, over convex alternatives such as the lasso. In recent years, several approaches to carrying out inference for lasso models have been developed, but there has been little work on inference for nonconvex penalties. Here, we discuss one approach based on local false discovery rates that can be extended to penalties such as MCP/SCAD in order to provide measures of significance for each feature selected by the model. This talk will also address the question of how inference for nonconvex penalties differs from that for lasso models from both a theoretical perspective and in simulations. Finally, we analyze data from a study using gene expression to predict survival in cancer patients to illustrate the insights that these methods provide in practice.

email: patrick-breheny@uiowa.edu

IT'S JUST A MATTER OF PERSPECTIVE - ROBUST REGRESSION FOR MICROBIOME DATA VIA PERSPECTIVE M-ESTIMATION

Christian L. Mueller*, Flatiron Institute, Simons Foundation

We recently introduced a model for maximum likelihood-type estimation (M-estimation) that generalizes a large class of existing statistical models, including Huber's concomitant M-estimation model and the scaled Lasso. The model, termed perspective M-estimation, leverages the observation that convex M-estimators with concomitant scale as well as various regularizers are instances of perspective functions. Such functions are amenable to proximal analysis, which leads to principled and provably convergent optimization algorithms via proximal splitting. In this contribution, we extend this framework to robust regression models with compositional covariates. Compositional (or relative abundance) data are commonplace in biology, including microbiome and metabolome measurements. We show attractive statistical performance of our new estimators on real-world compositional data and introduce non-convex extensions of our estimators that can handle heteroscedasticity in compositional data.

email: cmueller@flatironinstitute.org

RELAX AND SPLIT ALGORITHM FOR ICA

Peng Zheng*, University of Washington
Benjamin Risk, Emory University
Irina Gaynanova, Texas A&M University
Aleksandr Y. Aravkin, University of Washington

Independent component analysis (ICA) is an important unsupervised learning technique that is widely used in observational science including cognitive neuroscience and signal processing. It models the data as linear combination of independent non-Gaussian components which result the maximum likelihood formulation as a nonconvex optimization problem. In this work, we propse a "relax and split" optimization framework that address challenges raised by nonconvex ICA objective. We compare our method with existing ICA solver on both synthetic data and real fMRI data. Moreover, we incorporating new mechanism into ICA to encourage sparsity of the indenpent components. In the fMRI experiment, sparse ICA generate cleaner and more interpretable results that helps us to extract neuro activity signals; stablize the algorithm and make it more robust to the artifacts. Also, sparse ICA brings more reproducibility to the experiments compare usual approaches that require manually thresholding signal at the end.

email: zhengp@uw.edu

INTEGRATED PRINCIPAL COMPONENT ANALYSIS

Genevera I. Allen*, Rice University and Baylor College of Medicine
Tiffany M. Tang, University of California, Berkeley

The growth in data volume and variety drives the need for principled data integration methods that can analyze multiple sources of data simultaneously. To facilitate and improve such integrative data analyses, we develop a new statistical method, Integrated Principal Components Analysis (iPCA). iPCA is a generalization of the classical Principal Components Analysis that uses a Kronecker covariance model based on the matrix-variate normal distribution to capture individual patterns within each dataset and joint patterns shared by multiple data sets. We develop our iPCA model by proposing several classes of estimators, characterizing their optimization theoretic properties by showing that some achieve global optimality for a non-convex problem, and studying their statistical consistency. Our approach provides a firm model-based and theoretical foundation for dimension reduction, pattern recognition, visualization, and exploratory analysis of integrated data. We demonstrate the effectiveness of iPCAin simulations and real data examples, including an application to integrative genomics for Alzheimer's disease.

email: gallen@rice.edu

105. BIOPHARMACEUTICAL RESEARCH AND CLINICAL TRIALS

IN SILICO CLINICAL TRIAL SIMULATION AND VIRTUAL PATIENTS' GENERATION

Philippe Saint Pierre*, Toulouse Institute of Mathematics
Nicolas J. Savy, Toulouse Institute of Mathematics

In silico clinical trial (ISCT) have been recognized by the pharmaceutical companies and regulatory authorities as being crucial to improve the efficiency of the drug development process. This includes the use of ISCT to optimize trial designs at the various stages of development. An ISCT consist in studying a drug in a virtual patient population using available data and mathematical modeling. The generation of virtual patients is a key point of the process. The classical approaches to generate a multivariate database from a learning sample are based on Monte Carlo simulations from the joined distribution of the covariates (Discrete method) and on multi-normal distribution (Continuous method). These approaches present drawbacks when the dimension of covariates increased. We proposed a new approach for patients' generation based on R-vines copula. This kind of copula is particularly useful to deal with multiple correlations between variables. Modeling patients' evolution over time is another challenging task. Several execution models involving design study parameters and data-driven parameters can be proposed to study safety or efficiency of a drug.

email: Philippe.Saint-Pierre@math.univ-toulouse.fr

NONPARAMETRIC TESTS FOR TRANSITION PROBABILITIES IN MARKOV MULTI-STATE MODELS

Giorgos Bakoyannis*, Indiana University

Frequently, clinical trials involve event history data with multiple events. In such cases, the standard analysis methods compare hazard rates or transition intensities for particular events across treatment groups. However, these methods do not provide inference about transition probabilities, which directly reflect clinical prognosis. In this work, we propose nonparametric two-sample tests for transition probabilities in general nonhomogeneous Markov multi-state models. The asymptotic null distributions of the tests are derived and the tests are shown to be consistent against any alternative hypothesis. The latter implies that the tests provide good power for detecting a difference even between crossing transition probability curves. The proposed tests are also extended for studies where clustering is present, such as multicenter clinical trials and cluster-randomized trials. Simulation studies show good performance of the tests even with small sample sizes and under alternative hypotheses with crossing transition probability curves. Finally, the proposed tests are illustrated using real data from two randomized controlled trials.

email: gbakogia@iu.edu

TRIGGER STRATEGY IN REPEATED TESTS ON MULTIPLE HYPOTHESES

Jiangtao Gou*, Fox Chase Cancer Center, Temple University Health System

We proposed a trigger strategy as a general framework to extend a variety of single-stage designs to multistage adaptive designs. There are two types of triggers: significance trigger and time trigger. An important application is to hierarchically test multiple endpoints in a group sequential design. Popular hierarchical testing strategies include the stagewise hierarchical strategy, overall hierarchical strategy and partially hierarchical strategy. They are all special cases of trigger strategies. Meanwhile, trigger strategy provides many new flexible hierarchical testing rules and the corresponding methods to calculate the refined critical boundaries. By incorporating the information of correlation coefficients and hierarchical structures, the trigger strategy can suggest various alpha-exhausted procedures to boost the statistical power.

email: jiangtao.gou@fccc.edu

MCP-MOD FOR EXPOSURE-RESPONSE INFORMATION

Gustavo Amorim*, Vanderbilt University Medical Center
An Vandebosch, Janssen Pharmaceutica
Jose Pinheiro, Janssen Pharmaceutica
Joris Menten, Janssen Pharmaceutica
Kim Stuyckens, Janssen Pharmaceutica

Establishing the dose-response relationship between a compound and clinical endpoint is an important part of drug development. It is involved in most steps of clinical research and is central, in particular, for assessing the efficacy of the drug as well as for properly estimating the correct dose to be used in clinical practice. Establishing the dose-response profile has traditionally being addressed by analysis of dose-response data only. Recent studies, however, showed that more precise

estimates may be obtained if extra information from pharmacokinetics models are also used in addition to the dose levels under investigation. This gain in precision is achieved by shifting the focus from dose response to dose-exposure-response modeling, which uses inter-subject information to reduce response-uncertainty, which may translate into more accurate dose selections. In light of this, extensions of the well-known MCP-Mod procedure that fully uses the complete dose-exposure-information are proposed. Their performance are discussed via simulations studies and compared to standard dose-response analysis.

email: ggca@outlook.com

EVALUATING THE FINITE SAMPLE PROPERTIES OF BASELINE COVARIATE ADJUSTMENT IN RANDOMIZED TRIALS: APPLICATION TO TIME TO EVENT AND BINARY OUTCOMES

Su Jin Lim*, Johns Hopkins University School of Medicine
Elizabeth Colantuoni, Johns Hopkins Bloomberg School of Public Health

In randomized controlled trials (RCTs) with baseline covariates that are prognostic for the outcome of interest, baseline covariate adjustment can improve precision of the estimated marginal treatment effect. Baseline covariate adjusted estimators for the marginal treatment effect have been proposed for a variety of outcomes and the large sample properties of these estimators have been derived. However, there is little understanding of how baseline covariate adjusted estimators behave in small sample RCTs. We conduct extensive simulation studies evaluating the statistical behavior of adjusted estimators for both time to event and binary outcomes in small sample RCTs, defining the marginal treatment effect as the log hazard ratio and the absolute risk difference, respectively. Simulations demonstrate that precision gains can be achieved when adjusting for prognostic baseline covariates in small sample RCTs, but such gains depend on the strength of the correlation between the baseline covariates and the outcome, and the effective sample size. We discuss both potential precision gains and losses under varying correlation and sample size scenarios.

email: slim36@jhmi.edu

106. MISSING DATA

A DOUBLY-ROBUST METHOD TO HANDLE MISSING MULTILEVEL OUTCOME DATA WITH APPLICATION TO THE CHINA HEALTH AND NUTRITION SURVEY

Nicole M. Butera*, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
Annie Green Howard, University of North Carolina, Chapel Hill
Penny Gordon-Larsen, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill

Missing data are common in longitudinal cohort studies and can lead to bias, especially nonrandom missingness. Many common methods for handling nonrandom missing data require correctly specifying a missingness model. Although doubly robust methods exist to provide unbiased estimates in the presence of missing data, they do not handle correlation due to clustering from longitudinal or cluster-sampled studies. We developed a doubly robust method to estimate regression of an outcome on a predictor in the presence of missing multilevel outcome data, which results in consistent estimation of coefficients assuming correct specification of either (1) the missingness probability or (2) the outcome model. This method involves specification of multilevel models for missingness and the outcome, conditional on observed variables and cluster-specific random effects, to account for correlation. We showed this proposed estimator is doubly robust and asymptotically normal, conducted simulations to compare the method to an existing doubly robust method for independent data, and applied the method to data from the China Health and Nutrition Survey, an ongoing multilevel longitudinal cohort study.

email: butera@live.unc.edu

REPRODUCIBILITY OF HIGH THROUGHPUT EXPERIMENTS IN CASE OF MISSING DATA

Roopali Singh*, The Pennsylvania State University
Feipeng Zhang, The Pennsylvania State University
Qunhua Li, The Pennsylvania State University

The outcome of high-throughput biological experiments such as Single-cell RNA-seq, often has a lot of missing observations when the signals are below the detection level. Hence, understanding the effect of missing data on the estimation of reproducibility of the outcome is important. The existing methods for reproducibility assessment do not take account of the missing values, leading to biased results. In this paper, we study how the reproducibility of high-throughput experiments is affected by the choices of operational factors (e.g. platform or sequencing depth), when a large amount of measurements are missing. Using a latent variable approach, we extend the correspondence curve regression for reproducibility assessment, to incorporate missing values. In contrast to existing methods, our approach estimates the independent effects of covariates on reproducibility and the amount of missing data. Using simulations, we show that our method is more accurate in detecting difference in reproducibility than existing approaches. We illustrate the usefulness of our method using a study of HCT116 cells from scRNA-seq libraries made using microfluidic method and tube-based methods.

email: rus82@psu.edu

MULTIPLE IMPUTATION STRATEGIES FOR HANDLING MISSING DATA WHEN GENERALIZING RANDOMIZED CLINICAL TRIAL FINDINGS THROUGH PROPENSITY SCORE-BASED METHODOLOGIES

Albee Ling*, Stanford University
Maya Mathur, Stanford University
Kris Kapphahn, Stanford University
Maria Montez-Rath, Stanford University
Manisha Desai, Stanford University

Randomized clinical trials (RCTs) are the gold standard for estimating treatment effects, but the trial findings may not be applicable to target populations of interest. Propensity Score (PS)-based methods have been shown to mitigate this issue. Inverse probability of selection weighting (IPSW) can be used to reweight the trial sample as a function of the PS, allowing the trial sample to resemble the target population. Missing data in covariates used in the PS estimation can threaten the validity of such methods. Multiple Imputation (MI) is a well-established and accessible method for handling missing data, but there is no consensus on the best practice.

We conducted a simulation study to evaluate properties of estimators under a variety of MI strategies, coupled with ways to integrate PS into analyses. Using a real-world example of generalizing Frequent Hemodialysis Network (FHN) findings to United States Renal Data System (USRDS) data, we illustrate considerable heterogeneity across methods and provide practical guidelines.

email: yling@stanford.edu

PROPER SPECIFICATION OF MICE IMPUTATION MODELS FOR DATA WITH INTERACTIONS: NEW DEVELOPMENTS AND PRACTICAL RECOMMENDATIONS FOR R USERS

Emily Slade*, University of Kentucky

For implementing multiple imputation on data that are missing at random, it has been well-established that interactions of scientific interest must be captured in the imputation model to avoid bias in the estimated coefficients. Simulation work by Tilling et al. (2016) revealed this issue to be even more complex when employing multiple imputation by chained equations (MICE); namely, the presence of a single interaction term in the final analysis model requires an entire set of interactions to be present in the univariate imputation models. With this past work performed in Stata, R users may ask: based on these results, how should I specify MICE imputation models with interaction terms in R? The answer is nontrivial as multiple arguments in the mice function exist to accomplish the task, yet they differ in terms of bias and coverage of the estimated coefficients. Moreover, the question of whether passive imputation or just-another-variable imputation is more statistically valid for analyses with interaction terms still exists. We answer these questions via simulation and provide practical recommendations for the R user when performing MICE on data with interactions.

email: emily.slade@uky.edu

VARIANCE ESTIMATION WHEN COMBINING INVERSE PROBABILITY WEIGHTING AND MULTIPLE IMPUTATION IN ELECTRONIC HEALTH RECORDS-BASED RESEARCH

Tanayott Thaweethai*, Harvard University
Sebastien Haneuse, Harvard T.H. Chan School of Public Health
David Arterburn, Kaiser Permanente Washington Health Research Institute

Due to the complex process by which electronic health records (EHR) are generated and collected, missing data is a huge challenge when conducting large observational studies using EHR data. Most standard methods to adjust for selection bias due to missing data fail to address the heterogeneous structure of EHR data. Haneuse et al. (2016) proposed a method that considers a modularization of the data provenance, or the sequence of specific decisions or events that lead to observing complete data. Using this framework, a strategy has been developed that uses inverse probability weighting and multiple imputation at different stages to address missing data. We show that the proposed estimator is consistent and asymptotically Normal, and derive a consistent estimator of the asymptotic variance. A simulation study demonstrates these properties of the proposed estimator, and the variance estimator is compared with Rubin's standard combining rules for multiple imputation and a bootstrap-based imputation variance estimator.

email: tthaweethai@g.harvard.edu

MODEL-BASED PHENOTYPING IN ELECTRONIC HEALTH RECORDS WITH DATA FOR ANCHOR-LABELED CASES AND UNLABELED PATIENTS

Lingjiao Zhang*, University of Pennsylvania
Xiruo Ding, University of Pennsylvania
Yanyuan Ma, The Pennsylvania State University
Naveen Muthu, University of Pennsylvania
Jason Moore, University of Pennsylvania
Daniel Herman, University of Pennsylvania
Jinbo Chen, University of Pennsylvania

Building a classifier for a binary phenotype from Electronic Health Records normally requires a curated dataset consisting of both cases and controls. For some phenotypes, it is feasible to identify a group of cases upon specification of an anchor variable, but infeasible to identify control patients. An anchor variable being positive indicates cases, being negative is uninformative of the true phenotype status. We developed a likelihood approach to building a logistic model based classifier using both anchor-labeled cases and unlabeled patients. Our method spares researchers largely from labor-intensive manual labeling, leads to greatly increased efficiency. It yields consistent estimates for phenotype prevalence and anchor sensitivity. Additionally, we proposed novel statistical methods for assessing model calibration and predictive accuracy. We evaluated the performance of our method through theoretical and simulation studies, considering a range of phenotype prevalence and varying degree of model goodness-of-fit. We also applied the proposed method to identify patients with primary aldosteronism in the University of Pennsylvania Health System.

email: lingjiao@pennmedicine.upenn.edu

FULLY BAYESIAN IMPUTATION MODEL FOR NON-RANDOM MISSING DATA IN qPCR

Valeriia Sherina*, University of Rochester Medical Center
Matthew N. McCall, University of Rochester Medical Center
Tanzy M. T. Love, University of Rochester Medical Center

We propose a new statistical approach to obtain differential gene expression of non-detects in quantitative real-time PCR (qPCR) experiments through Bayesian hierarchical modeling. We propose to treat non-detects as non-random missing data, model the missing data mechanism, and use this model to impute Ct values or obtain direct estimates of relevant model parameters. A typical laboratory does not have the resources to perform experiments with a large number of replicates; therefore, we propose an approach that does not rely on large sample theory. We aim to demonstrate the possibilities that exist for analyzing qPCR data in the presence of non-random missingness through the use of Bayesian estimation. In this work we introduce and describe our hierarchical model and chosen prior distributions, assess the model sensitivity to the choice of prior, perform convergence diagnostics for the Markov Chain Monte Carlo, and present the results of a real data application.

email: valery.sherina@gmail.com

107. BAYESIAN COMPUTATIONAL AND MODELING METHODS

SAMPLING PRUDENTLY USING INVERSION SPHERES (SPInS) ON THE SIMPLEX

Sharang Chaudhry*, University of Nevada Las Vegas
Daniel Lautzenheiser, University of Nevada Las Vegas
Kaushik Ghosh, University of Nevada Las Vegas

In the computational Bayesian paradigm, optimal sampling can often be a challenging task. The complexity of the problem is exacerbated when the parameters of interest have external constraints like the sum-to-one, which naturally occurs in various fields such as compositional data analysis, signal separation, and neuroimaging. The sum-to-one constraint can be equivalently thought of as the parameters being constrained within the probability simplex. In this work, a transformation called inversion-in-a-sphere is introduced within the Metropolis-Hastings algorithm to make the simplex amenable to sampling. The performance of the proposed method is evaluated using simulation studies and comparative analyses. Application of this procedure is demonstrated using an example from neuroimaging.

email: sharang.chaudhry@unlv.edu

BAYESMETAB: BAYESIAN MODELLING APPROACH IN TREATING MISSING VALUES IN METABOLOMIC STUDIES

Jasmit Shah*, Aga Khan University
Guy N. Brock, The Ohio State University
Jeremy Gaskins, University of Louisville

With the rise of metabolomics, the development of methods to address analytical challenges in the analysis of metabolomics data is of great importance. Missing values (MVs) are pervasive, yet the treatment of MVs can have a substantial impact on downstream statistical analyses. The MVs problem in metabolomics is quite challenging, and can arise because the metabolite is not biologically present in the sample, or is present in the sample but at a concentration below the lower limit of detection (LOD), or is present in the sample but undetected due to technical issues related to sample pre-processing steps. In this study we propose a Bayesian modeling approach called BAYESMETAB to feature a cohesive and robust modeling structure for MVs in high dimensional metabolomics data. Our model accounts for MVs due to the truncation threshold, as well as other sources of missingness unrelated to true metabolite abundance. Statistical inference and data imputation are performed simultaneously using an MCMC algorithm. A hypothesis testing framework for differential abundance of metabolites between treatment group is considered, and BAYESMETAB is shown to perform better.

email: jasmit.shah@louisville.edu

A BAYESIAN MARKOV MODEL FOR PERSONALIZED BENEFIT-RISK ASSESSMENT

Dongyan Yan*, University of Missouri
Subharup Guha, University of Florida
Chul Ahn, U.S. Food and Drug Administration
Ram Tiwari, U.S. Food and Drug Administration

The development of systematic and structured approaches to assess benefit-risk of medical products is a major challenge for regulatory decision makers. In this article, we propose a Bayesian Markov model that treats the withdrawal category as an absorbing state, and analyze the subject-level data with multiple visits. A Log-odds ratio model is used to model the subject-level effects, by assuming a ratio of transition probabilities, with respect to a "reference" probability. A Dirichlet process is used as a prior for the subject-level effects to capture the similarity among subject response profiles. We develop an efficient Markov chain Monte Carlo algorithm for implementing the proposed method, and illustrate the estimation of individual benefit-risk profiles through simulation. The model's performance is evaluated using two model selection approaches, namely, the deviance information criterion (DIC) and the log-pseudo marginal likelihood (LPML). We analyze a clinical trial data using the proposed method to assess the subject-level or personalized benefit-risk in each arm, and to evaluate the aggregated benefit-risk difference between the treatments.

email: dyyr2@mail.missouri.edu

ITERATED MULTI-SOURCE EXCHANGEABILITY MODELS FOR INDIVIDUALIZED INFERENCE WITH AN APPLICATION TO MOBILE SENSOR DATA

Roland Z. Brown*, University of Minnesota
Julian Wolfson, University of Minnesota

Researchers are increasingly interested in using sensor technology to collect accurate activity information and make individualized inference about treatments, exposures, and policies. How to optimally combine population data with data from an individual remains an open question. Multi-source exchangeability models (MEMs) are a Bayesian approach for increasing precision by combining potentially heterogeneous supplemental data sources into analysis of a primary source. MEMs are a potentially powerful tool for individualized inference but can integrate only a few sources; their model space grows exponentially, making them intractable for high-dimensional applications. We propose iterated MEMs (iMEMs), which identify a subset of the most exchangeable sources prior to fitting a MEM model. iMEM complexity scales linearly with the number of sources, and iMEMs greatly increase precision while maintaining desirable asymptotic and small sample properties. We apply iMEMs to individual-level behavior and emotion data from a smartphone app and show that they achieve individualized inference with up to 99% efficiency gain relative to standard analyses that do not borrow information.

email: brow4288@umn.edu

A NOVEL BAYESIAN PREDICTIVE MODELLING IN TIME-TO-EVENT ANALYSIS USING MULTIPLE-IMPUTATION TECHNIQUES

Zhe (Vincent) Chen*, Novartis Pharmaceuticals Corporation
Kalyanee Viraswami-Appanna, Novartis Pharmaceuticals Corporation

Consider an oncology clinical trial with a time-to-event endpoint (for example, progression-free survival) in which an interim analysis has been performed. The time-to-event endpoint of interest is analyzed using standard Kaplan-Meier methodology and point estimates along with 95% confidence interval of the median time and event-free probability are reported. Given that there may be many patients censored at the time of the interim analysis, as the trial is ongoing, it is of interest to predict the

number of events and impact on the parameter estimates at the time of next analysis. In this work, we developed a statistical methodology, based on a combination of Bayesian and multiple-imputation techniques, which can be used to predict a wide range of parameter estimates after additional periods of study follow-up. We implement our methodology on synthetic data simulated under the underlying model, and show good prediction results.

email: vincent.chen@novartis.com

A LATENT CLASS BASED JOINT MODEL FOR RECURRENCE AND TERMINATION: A BAYESIAN RECOURSE

Zhixing Xu*, Florida State University
Debjayoti Sinha, Florida State University
Jonathan Bradley, Florida State University

In many clinical studies, each patient is at risk of recurrent events as well as the terminating event. We present a novel latent-class based semiparametric joint model that offers clinically meaningful and estimable association between the recurrence profile and risk of termination. Unlike previous shared-frailty based joint models, this model has coherent interpretation of the covariate effects on all relevant functions and model quantities that are either conditional or unconditional on events history. We offer a fully Bayesian method for estimation and prediction using a complete specification of the prior process of the baseline functions. When there is a lack of prior information about the baseline functions, we derive a practical and theoretically justifiable partial likelihood based semiparametric Bayesian approach. Our Markov Chain Monte Carlo tools for both Bayesian methods are implementable via publicly available software. Practical advantages of our methods are illustrated via a simulation study and the analysis of a transplant study with recurrent Non-Fatal Graft Rejections (NFGR) and the termination event of death due to total graft rejection.

email: zhixing.bruce.xu@gmail.com

108. CAUSAL EFFECT MODELING (MEDIATION/VARIABLE SELECTION/LONGITUDINAL)

THE ROLE OF BODY MASS INDEX AT DIAGNOSIS ON BLACK-WHITE DISPARITIES IN COLORECTAL CANCER SURVIVAL: A DENSITY REGRESSION MEDIATION APPROACH

Katrina L. Devick*, Harvard T.H. Chan School of Public Health
Linda Valeri, Columbia Mailman School of Public Health
Jarvis Chen, Harvard T.H. Chan School of Public Health
Alejandro Jara, Pontificia Universidad Católica de Chile
Marie-Abèle Bind, Harvard University
Brent A. Coull, Harvard T.H. Chan School of Public Health

The study of racial/ethnic inequalities in health is important to reduce the uneven burden of disease. In the case of colorectal cancer (CRC), disparities in survival among non-Hispanic Whites and Blacks are well documented, and mechanisms leading to these disparities need to be studied formally. Body mass index (BMI) is a well-established risk factor for developing CRC, and recent literature shows BMI at diagnosis of CRC is associated with survival. Since BMI varies by racial/ethnic group, a question that arises is whether disparities in BMI is partially responsible for observed racial/ethnic disparities in CRC survival. This paper presents new methodology to quantify the impact of the hypothetical intervention that matches the BMI distribution in the Black population to a potentially complex distributional form observed in the White population on racial/ethnic disparities in survival. We perform a simulation that shows our proposed Bayesian density regression approach performs as well as or better than current methodology allowing for a shift in the mean of the distribution only, and that standard practice of categorizing BMI leads to large biases.

email: kdevick@hsph.harvard.edu

UNIFIED MEDIATION ANALYSIS APPROACH TO COMPLEX DATA OF MIXED TYPES VIA COPULA MODELS

Wei Hao*, University of Michigan
Peter X.K. Song, University of Michigan

Motivated by pervasive biomedical data, we propose a unified mediation analysis approach to complex data of mixed types, including continuous, categorical, count variables. We invoke copula models to specify joint distributions of outcome variables, mediators and exposure variables of interest in the context of generalized linear models. We develop inference procedures to evaluate casual pathways in both aspects of parameter estimation and hypothesis testing for direct and/or indirect effects of the exposure variable on outcome variables. Our proposed method also enables us to identify important mediators through which exposure variables have indirect effects. We examine necessary model assumptions for the identifiability of casual effects and establish asymptotic properties for the proposed method. We compare the performance of the proposed method with other existing methods using simulation studies. We apply the proposed method to an analysis of a real biomedical dataset.

email: weihao@umich.edu

ESTIMATING CAUSAL MEDIATION EFFECTS FROM A SINGLE REGRESSION MODEL

Christina T. Saunders*, Vanderbilt University
Jeffrey D. Blume, Vanderbilt University

I will present a regression framework for estimating causal mediation effects (and their variance) from the fit of a single regression model, rather than from a system of equations (Saunders and Blume, Biostatistics). Requiring the fit of only one model permits the use of a rich suite of regression tools that are not easily implemented on a system of equations. I will highlight situations in which the difference and product of coefficients approaches yield different estimates of the total effect, even for linear models, and show how our new approach addresses this issue. I provide examples from complex research hypotheses, including models with multiple mediators, interactions, and nonlinearities. Using an example from genetic epidemiology, we compare our approach to existing methods such as the mediation formula and the KHB method. For large datasets (such as GWAS data), the proposed single-model framework also imparts substantial gains in computational efficiency.

email: christinatrippsaunders@gmail.com

MEDIATOR SELECTION VIA THE LASSO WITH NONPARAMETRIC CONFOUNDING CONTROL

Jeremiah Jones*, University of Rochester
Ashkan Ertefaie, University of Rochester

Researchers are often interested in learning not only the effect of treatments on outcomes, but the pathways through which these effects operate. Mediation analysis seeks to provide information on these pathways. A mediator is a variable that is affected by treatment and subsequently affects outcome. When few mediators are considered and the model is correctly specified, methods exist to estimate mediation effects. However, in practice linear associations are often posited. If confounders are associated nonlinearly with treatment, outcome, or the mediator variables, misspecification may bias the effect estimates. We propose to reduce the chance of model misspecification by estimating confounding effects nonparametrically. Even in the absence of misspecification, many methods for mediation analysis are intractable when the number of mediators is larger than the sample size. We demonstrate inflated variance from including noise variables in the mediation pathway. We propose a variable selection technique which extends the Adaptive Lasso to this setting using customized weights. The performance of these methods will be discussed and demonstrated through simulation studies.

email: jeremiah_jones@urmc.rochester.edu

ESTIMATING TIME-VARYING CAUSAL EFFECT MODERATION IN MOBILE HEALTH WITH BINARY OUTCOMES

Tianchen Qian*, Harvard University
Hyesun Yoo, University of Michigan
Predrag Klasnja, University of Michigan
Daniel Almirall, University of Michigan
Susan A. Murphy, Harvard University

Binary outcome is common in mobile health studies. We focus on estimating the time-varying causal effect moderation for data from micro-randomized trials with binary outcomes. We give the definition of moderated treatment effect in this setting, and provide two estimation methods. One estimation method is for the proximal treatment effect conditional on the entire history, and the estimator is semiparametric locally efficient. The other estimation method, based on weighted and centered least squares, is for the proximal treatment effect marginal but conditional only on a subset of variables in history. Both estimators are robust in the sense that they do not require a correct model for the outcome process. The methods are illustrated by simulation studies and a data example using HeartSteps data set, a mobile health study for increasing physical activity in participants.

email: qiantianchen@fas.harvard.edu

ESTIMATING CAUSAL EFFECTS WITH LONGITUDINAL DATA IN A BAYESIAN FRAMEWORK

Kuan Liu*, University of Toronto
Olli Saarela, University of Toronto
Eleanor Pullenayegum, University of Toronto, The Hospital for Sick Children

Causal inference methods to control confounding bias of observational study, and to accommodate time-dependent data structure of repeated measures, are of high interest. Advantages of Bayesian formulation include the propagation of propensity score estimation uncertainty, probabilistic summaries, and the flexibility of incorporating prior clinical beliefs. Despite recent interest in Bayesian causal methods, limited literature explored approaches to handle longitudinal data and none with repeatedly measured outcome. In this paper, we extended two Bayesian approaches, the Bayesian estimation of marginal structural models and the two-stage Bayesian propensity score analysis with explicitly defined repeatedly measured outcome. Our proposed methods permit causal estimation of treatment effects at each visit. Time-dependent propensity scores and inverse probability of treatment weights are obtained from the MCMC samples of the posterior treatment assignment model at each follow-up visit. We used a simulation study to validate and compare the proposed methods and illustrated our approaches through an efficacy study of intravenous immunoglobulin therapy in juvenile dermatomyositis.

email: kuan.liu@mail.utoronto.ca

BRAND VS. GENERIC: ADDRESSING NON-ADHERENCE, SECULAR TRENDS, AND NON-OVERLAP

Lamar Hunt*, Johns Hopkins Bloomberg School of Public Health and OptumLabs Visiting Fellows
Irene B. Murimi, Johns Hopkins Bloomberg School of Public Health and OptumLabs Visiting Fellows
Daniel O. Scharfstein, Johns Hopkins Bloomberg School of Public Health
Jodi B. Segal, Johns Hopkins School of Medicine
Marissa J. Seamans, Johns Hopkins Bloomberg School of Public Health
Ravi Varadhan, Johns Hopkins Center on Aging and Health

While generic drugs offer a cost-effective alternative to brand, regulators need a method to assess therapeutic equivalence in a post market setting when concerns about generic drug performance arise. This requires identifying causal effects in the presence of non-adherence, secular trends, and treatment non-overlap due to sharp uptake of the generic once it is available (a positivity violation). We identify a causal effect in a survival analysis setting by extending regression discontinuity to survival curves that are identified using G-Computation. Using insurance claims provided by OptumLabs®, we apply the method to immediate release venlafaxine. The sample includes 42,372 patients initiating between 1994 and 2016. Failure is defined as a composite of (1) treatment change (treatment with an anti-depressant other than venlafaxine, an anti-psychotic, lithium, or electroconvulsive therapy), (2) clinical progression (suicide related clinical encounter, hospitalization or emergency department visit), and (3) death. The restricted mean difference in survival is -0.18 days with a standard error of 13.32. We find no evidence for a difference in therapeutic equivalence.

email: lhunt13@jhmi.edu

109. MICROBIOME DATA ANALYSIS WITH ZERO INFLATION AND/OR MODEL SELECTION

AN INTEGRATIVE BAYESIAN ZERO-INFLATED NEGATIVE BINOMIAL MODEL FOR MICROBIOME DATA ANALYSIS

Shuang Jiang*, Southern Methodist University
Guanghua Xiao, University of Texas Southwestern Medical Center
Andrew Y. Koh, University of Texas Southwestern Medical Center
Yang Xie, University of Texas Southwestern Medical Center
Qiwei Li, University of Texas Southwestern Medical Center
Xiaowei Zhan, University of Texas Southwestern Medical Center

In this paper, we developed an integrative Bayesian hierarchical mixture model to analyze the human microbiome data for multiple patient groups. The observed microbial abundances are over-dispersed count data with a large number of zeros, and the count of each taxon is potentially affected by both patient's disease type and genetic covariates. Based on a zero-inflated negative binomial framework, we aim to distinguish a subset of taxa that best discriminate between patient groups and to simultaneously incorporate biological confounders to reduce false positive findings. Our extensive simulations show that our approach has favorable performance in feature selection compared with other commonly used methods. In the analysis of real microbiome datasets, our method detects microbiota associated with patients' conditions and selects significant associations between microbiota and metabolic pathways or metabolites. Biological interpretations of our results confirm those of previous studies and offer a more comprehensive understanding of the underlying mechanism in disease etiology.

email: shuangj@smu.edu

BAYESIAN HIERARCHICAL ZERO-INFLATED NEGATIVE BINOMIAL MODELS WITH APPLICATIONS TO HIGH-DIMENSIONAL HUMAN MICROBIOME COUNT DATA

Amanda H. Pendegraft*, University of Alabama at Birmingham
Nengjun Yi, University of Alabama at Birmingham

Zero-inflation is a characteristic of data obtained via classification of microbial 16S rRNA genes extracted from human tissues. In other words, human microbiome count data contains a large number of features not frequently observed in a large number of samples. Reasons for sparsity are two-fold: a feature may be undetectable at a pre-specified depth of coverage or a feature may simply not present. Zero-inflated models have been proposed to delineate between excessive zeros, however, techniques are needed to concurrently adjust for high-dimensional combinations involving tens or hundreds of potentially correlated covariates expanding applications of human microbiome count data. A Bayesian hierarchical zero-inflated negative binomial model is introduced to control for complex high-dimensional study designs utilizing zero-inflated data. An application is considered using subsets of the American Gut project participants diagnosed by a medical professional with inflammatory bowel disease and/or irritable bowel syndrome following propensity score matching to controls.

email: alhall91@uab.edu

MODEL SELECTION FOR LONGITUDINAL MICROBIOME DATA WITH EXCESS ZEROS

Tony A. Chen*, Princeton University
Yilun Sun, St. Jude Children's Research Hospital
Hana Hakim, St. Jude Children's Research Hospital
Ronald Dallas, St. Jude Children's Research Hospital
Jason Rosch, St. Jude Children's Research Hospital
Sima Jeha, St. Jude Children's Research Hospital
Li Tang, St. Jude Children's Research Hospital

Despite the rich body of statistical literature on models to analyze data with excess zeros, overdispersed counts, and repeated measures, it is unclear how to evaluate and compare the performance of such models with microbiome data. Within a parametric framework, our study evaluates the reliability of common model selection criteria for generalized linear mixed models for longitudinal microbiome data. Simulations of microbiome data show that criteria such as AIC can select the appropriate model for given data. In the case of zero-inflated data with low overdispersion, the Negative Binomial mixed model (NBMM) was often chosen over Zero-Inflated and Hurdle NBMMs, suggesting that a zero-inflation component may only be necessary when the data are also overdispersed. When analyzing a study cohort of acute lymphoblastic leukemia patients at St. Jude Children's Research Hospital, the traditional NBMM was favored most often, which supports its applicability to real microbiome data. Overall, we find common model selection criteria appropriate in longitudinal microbiome data analysis, though further simulation studies are necessary to fully understand the behavior of these models.

email: tonyac@princeton.edu

BAYESIAN VARIABLE SELECTION IN REGRESSION WITH COMPOSITIONAL COVARIATES

Liangliang Zhang*, University of Texas MD Anderson Cancer Center

The microbiome data are very high dimensional, as the number of OTUs is large and these OTU abundances are usually aggregated to different taxonomical levels. In this paper, we consider regression analysis with microbiome compositional covariates. Our goal is to identify the bacterial taxa that are associated with a continuous response such as the body mass index (BMI). A phylogenetic tree can be viewed as an undirected graph, providing us prior knowledge of associations and grouping effects of microbiome compositional covariates. We approached this problem through Bayesian variable selection framework by considering the phylogenetic tree as a graphical prior and formulated an Ising prior on the model space to incorporate structural information. We specialized and generalized the model for microbiome compositional data which includes the log-transform of data to avoid regulation with linear constraint, quantifying the tree structure to graphical associations and validating the performance of our model.

email: lions_z@hotmail.com

COMPOSITIONAL KNOCKOFF FILTER FOR FDR CONTROL IN MICROBIOME REGRESSION ANALYSIS

Arun A. Srinivasan*, The Pennsylvania State University
Lingzhou Xue, The Pennsylvania State University
Xiang Zhan, The Pennsylvania State University

A critical task in microbiome analysis is to identify microbial taxa that are associated with a response of interest. Many statistical methods examine the association between the response and one microbiome feature at a time, followed by multiple testing adjustments such as false discovery rate (FDR) control. These methods are often underpowered due to the properties of microbiome data, such as high-dimensionality and a compositional constraint. We propose the compositional knockoff filter (CKF) to provide finite-sample FDR control using linear log-contrast models for regression analysis of microbiome compositional data. Our proposed compositional knockoff filter achieved the FDR control in a regression model that jointly analyzes the microbiome community. Firstly, we introduce the use of Constrained Best Subset Selection to reduce the high-dimensional model to a tractable low-dimensional setting. Secondly, we impose a l1-regularization in the regression model. A subset of the microbes are selected under a pre-specified FDR threshold. The method is demonstrated via simulation studies, and an application to a microbiome study.

email: uus91@psu.edu

GENERALIZED BIPLOTS FOR THE ANALYSIS OF HUMAN MICROBIOME

Yue Wang*, Fred Hutchinson Cancer Research Center
Timothy W. Randolph, Fred Hutchinson Cancer Research Center
Ali Shojaie, University of Washington
Jing Ma, Fred Hutchinson Cancer Research Center

In the analysis of human microbiome, dimension-reduced graphical displays often reveal meaningful microbial community structure across human subjects. These ordination methods are based on biologically defined similarity. For example, principal coordinates analysis (PCoA) begins with a matrix of pairwise distances or dissimilarities between vectors of taxon abundances, which allows analyses to incorporate non-Euclidean structure. However, once the distance is calculated, it is no longer clear which taxon is important to the observed clustering. We first construct approximate decompositions of the data matrix based on generalized matrix decomposition (GMD) and show how the decompositions can be used to construct contribution biplots, which include two sets of points corresponding to human subjects and taxa respectively. Furthermore, based on the method of kernel-penalized regression, we allow the biplot to only include the set of taxa which are significantly associated with a given outcome or phenotype. We illustrate our approach using two recent studies on gut and vaginal microbiomes.

email: ywang2310@fredhutch.org

110. RECURRENT EVENTS OR MULTIPLE TIME-TO-EVENT DATA

REGRESSION ANALYSIS OF RECURRENT EVENT DATA WITH MEASUREMENT ERROR

Yixin Ren*, University of Maryland, College Park
Xin He, University of Maryland, College Park

This paper considers regression analysis of recurrent event data in the presence of measurement error in covariates. We present a class of semiparametric models for recurrent events that allows correlations between censoring times and recurrent event process via frailty. Estimating equations are developed for estimation of regression parameters, and both large and finite sample properties of the proposed estimates are established. An illustrative example from a clinical trial is provided.

email: yr0403@math.umd.edu

A GENERAL CLASS OF SEMIPARAMETRIC MODELS FOR BIASED RECURRENT EVENT DATA

Russell S. Stocker*, Indiana University of Pennsylvania
Akim Adekpedjou, Missouri University of Science and Technology

We propose a general class of weighted semiparametric models for recurrent event data that constitute a biased sample of the target population. To correct for the bias, estimated weights obtained via logisitic regression are assigned to each unit. A weighted class of semiparametric models is then fitted to the data. We use the tools of empirical process theory to establish the asymptotic properties of the estimators. A computer simulation study indicates that the finite sample properties of the estimators are well approximated by their asymptotic properties. A real data set is analyzed to illustrate the class of models.

email: rstocker@iup.edu

PENALIZED SURVIVAL MODELS FOR THE ANALYSIS OF ALTERNATING RECURRENT EVENT DATA

Lili Wang*, University of Michigan
Kevin He, University of Michigan
Douglas E. Schaubel, University of Michigan

Recurrent event data are widely encountered in clinical and observational studies. Most methods for recurrent events treat the outcome as a point process and, as such, neglect any associated event duration. This generally leads to a less informative and potentially biased analysis. We propose a joint model for the recurrent event rate (of incidence) and duration. The two processes are linked through a bivariate normal frailty. For example, when the event is hospitalization, we can treat the time to admission and length-of-stay as two alternating recurrent events. In our method, the regression parameters are estimated through a penalized partial likelihood, and the variance-covariance matrix of the frailty is estimated through a recursive estimating formula. Simulation results demonstrate that our method provides accurate

parameter estimation, with relatively fast computation time. We illustrate the methods through an analysis of hospitalizations among end-stage renal disease patients.

email: lilywang@umich.edu

A TIME-VARYING JOINT FRAILTY-COPULA MODEL FOR ANALYZING RECURRENT EVENTS AND A TERMINAL EVENT: AN APPLICATION TO THE CARDIOVASCULAR HEALTH STUDY

Zheng Li*, Novartis Pharmaceuticals Corporation
Vernon M. Chinchilli, The Pennsylvania State University
Ming Wang, The Pennsylvania State University

Recurrent events could be stopped by a terminal event, which commonly occurs in biomedical and clinical studies. In this situation, the non-informative censoring assumption could be violated. The joint frailty model is widely used to jointly model these two processes. However, several limitations exist: 1) recurrent events and terminal event processes are conditionally independent given the subject-level frailty and 2) the correlation between the terminal event and the recurrent events is constant over time. We propose a time-varying joint frailty-copula model to relax these two assumptions in the Bayesian framework. Conditional on the frailty, the survival functions are jointly modeled by a survival copula and the dynamic correlation between the terminal event and the recurrent event process is modeled by a latent Gaussian AR(1) process. The simulation results show that compared with the joint frailty model and the joint frailty-copula model, the absolute bias and mean squared error of the proposed method is the smallest. We applied our method to analyze the CHS data to identify risk factors to myocardial infarctions or strokes, and quantify their correlation with death.

email: zheng.li@novartis.com

DYNAMIC REGRESSION WITH RECURRENT EVENTS

Jae Eui Soh*, Emory University
Yijian Huang, Emory University

Recurrent events often arise in follow-up studies where a subject may experience multiple occurrences of the same event. Most regression models with recurrent events tacitly assume constant effects of covariates. However, the effects may actually vary over time in many applications. To address the time-varying effects, we propose a dynamic regression model to target the mean frequency of recurrent events at the population level. We develop an estimation procedure that fully exploits the observed data. Consistency and weak convergence of the proposed estimator are established. Simulation studies demonstrate that the proposed method works well, and two real data analyses illustrate its practicality.

email: statsoh@gmail.com

SPEARMAN'S CORRELATION FOR ESTIMATING THE ASSOCIATION BETWEEN TWO TIME-TO-EVENT OUTCOMES

Svetlana K. Eden*, Vanderbilt University
Chun Li, Case Western Reserve University
Bryan E. Shepherd, Vanderbilt University

There is often interest in measuring association between two time-to-event variables measured on one or different individuals, e.g., times to events in twins, or times to viral failure and regimen change among persons being treated for HIV. The problem is complicated due to right censoring. Spearman-like correlation measures have been proposed, but they require parametric assumptions that are incompatible with the non-parametric spirit of Spearman's correlation. We propose two new non-parametric approaches. Our first method is an extension of Spearman's correlation for censored data using probability-scale residuals that requires estimating only marginal survival curves, and is related to a classical test of association for bivariate survival data. Our second method computes Spearman's rank correlation using a non-parametric estimator of the bivariate survival surface. We also propose restricted versions of these estimators for settings where survival surfaces are not estimable past certain time. We study properties of our estimators, and we demonstrate their performance using simulations. We apply our methods to an HIV cohort initiating antiretroviral therapy in Latin America.

email: svetlana.eden@Vanderbilt.Edu

111. INDIVIDUALIZED EVIDENCE FOR MEDICAL DECISION MAKING: PRINCIPLES AND PRACTICES

BAYESIAN HIERARCHICAL MODELS FOR INDIVIDUALIZED HEALTH

Scott L. Zeger*, Johns Hopkins University

This talk will frame key health questions in statistical terms, then offer a Bayesian hierarchical modeling approach that integrates potentially complex data with prior biomedical knowledge to empirically address the questions. The talk will present two applications, the first to improve clinical decisions about prostate cancer care, and the second to better understand the etiology of autoimmune disease. It will identify opportunities for statistical scientists to directly impact health decisions by translating their models into practice for use locally and as a component of information networks of the future. It will identify obstacles to be overcome by teams on which statisticians will play key roles.

email: sz@jhu.edu

PERSONALIZED BAYESIAN MINIMUM-RISK DECISIONS FOR TREATMENT OF CORONARY ARTERY DISEASE

Laura A. Hatfield*, Harvard Medical School

Bayesian decision theory establishes that the optimal statistical decision minimizes risk, that is, the loss function integrated over the posterior of the unknown variables. We have previously extended this framework to real-world treatment decisions using stylized loss functions that combine expected treatment outcomes with patient

preferences (in the form of health utilities). That preference-weighted outcome score (PWOS) method incorporated uncertainty in expected outcomes but included only binary outcomes. In the current work, we extend the method to time-to-event outcomes, accounting for semi-competing risks and patient covariates. We use simulation to show the potential to improve decisions compared to methods that treat patient preferences as fixed or ignore uncertainty in outcomes of treatment. We demonstrate the potential of this method by applying it to the decision among drug-eluting stents to treat coronary artery disease.

email: hatfield@hcp.med.harvard.edu

ASSESSING POTENTIAL CLINICAL IMPACT WITH NET BENEFIT MEASURES

Tracey L. Marsh*, Fred Hutchinson Cancer Research Center

Referral strategies based on risk scores and medical tests provide one approach to individualizing clinical care. Direct assessment of clinical impact requires implementing the strategy and is not possible in early phases of biomarker research. However, the potential clinical impact of a proposed strategy can be assessed by using a net benefit analysis. Net benefit combines sensitivity and specificity into a single population-level measure that can be calculated from early-phase retrospective studies. Introduced in the context of graphical tools for evaluating risk prediction models, net benefit measures have also been used to define which sensitivity and specificity values might confer clinical utility and to set performance targets in biomarker discovery. After introducing net benefit in the setting of a validation study, I will highlight recent work on statistical inference for net benefit. I will motivate and demonstrate the methodology with examples from cancer biomarker research and emergency medicine.

email: tmarsh@fredhutch.org

INDIVIDUALIZED EVIDENCE FOR MEDICAL DECISION MAKING: PRINCIPLES AND PRACTICES

David M. Kent*, Tufts Medical Center

A fundamental incongruity in evidence-based medicine (EBM) is that evidence is derived from groups of people yet medical decisions are made for individuals. Popular approaches to EBM have encouraged the direct application of average effects estimated in clinical trials to guide decision making for individuals, as though all patients meeting trial inclusion criteria are likely to experience similar benefits from treatments. This attitude has proven remarkably durable and compelling, despite variation in patient characteristics and outcomes seen in clinical practice. Conventional ("one-variable-at-a-time") subgroup analysis ignores the fact that patients have multiple characteristics simultaneously that affect the likelihood of treatment benefit. These analyses are also typically underpowered and vulnerable to spurious results from multiplicity. Using clinical examples, I will demonstrate the potential benefits of using more comprehensive subgrouping schemes that incorporate information on multiple variables, such as those based on summary variables (e.g., risk scores) and review new recommendations for "predictive approaches" to analysis of heterogeneous treatment effect.

email: dkent1@tuftsmedicalcenter.org

112. SOME NEW PERSPECTIVES AND DEVELOPMENTS FOR DATA INTEGRATION IN THE ERA OF DATA SCIENCE

USING SYNTHETIC DATA TO UPDATE AN ESTABLISHED PREDICTION MODEL WITH NEW BIOMARKERS

Jeremy Taylor*, University of Michigan
Tian Gu, University of Michigan
Bhramar Mukherjee, University of Michigan

We consider the situation where there is an established regression model that can be used to predict an outcome, Y, from predictor variables X. A new variable B may enhance the prediction of Y. A dataset of size n containing Y, X and B is available, and the challenge is to build a model for [Y|X,B] that uses both the available data and the known model for [Y|X]. We propose a synthetic data approach that consists of creating m additional synthetic data observations, and then analyzing the combined dataset of size n+m to estimate the [Y|X,B] model. The synthetic data is created by replicating X then generating a synthetic value of Y from the known [Y|X] distribution. This combined dataset has missing values of B for m of the observations, and is analyzed using methods that can handle missing data. In special cases when [Y,X,B] is trivariate normal or when all of Y,X and B are binary we show that the synthetic data approach with very large m gives identical asymptotic variance for the parameters of the [Y|X,B] model as a constrained maximum likelihood estimation approach. This provides some theoretical justification and given its broad applicability makes the approach very appealing.

email: jmgt@umich.edu

INTEGRATIVE DATA ANALYTICS AND CONFEDERATE INFERENCE

Peter XK Song*, University of Michigan
Lu Tang, University of Pittsburgh
Ling Zhou, University of Michigan

This talk concerns integrative data analytics and statistical algorithms in data integration. As data sharing from related studies become of interest, statistical methods for a joint analysis of all available datasets are needed in practice to achieve better statistical power and detect signals that are otherwise impossible based on a single dataset alone. A major challenge arising from integrative data analytics pertains to principles of information aggregation, learning data heterogeneity, algorithms for model fusion. Information aggregation has been studied extensively by many statistics pioneers, which lay down the foundation of data integration. In this process, it is of critical importance to accommodate data heterogeneity, and otherwise the analysis will result in biased estimation and misleading inference. Distributed computing and confederate inference will be discussed, which enable to overcome some significance barriers in data integration.

email: pxsong@umich.edu

GENERALIZED META-ANALYSIS FOR DATA INTEGRATION WITH SUMMARY-LEVEL DATA

Nilanjan Chatterjee*, Johns Hopkins University
Prosenjit Kundu, Johns Hopkins University
Runlong Tang, Johns Hopkins University

We propose developing a generalized meta-analysis (GENMETA) approach for combining information on multivariate regression parameters across multiple different studies which have varying 20 level of covariate information. Using algebraic relationships between regression parameters in different dimensions, we specify a set of moment equations for estimating parameters of a maximal model through information available from sets of parameter estimates from a series of reduced models. The specification of the equations requires a reference dataset to estimate the joint distribution of the covariates. We propose to solve these equations using the generalized method of moments approach, with the optimal weighting of the equations taking into account uncertainty associated with estimates of the parameters of the reduced models. Based on the same moment equations, we also propose a diagnostic test for detecting violation of underlying model assumptions, such as those arising due to heterogeneity in the underlying study populations. Methods are illustrated using extensive simulation studies and a real data example involving the development of a breast cancer risk prediction model.

email: nilanjan10c@jhu.edu

A SUPERPOPULATION APPROACH TO CASE-CONTROL STUDIES

Yanyuan Ma*, The Pennsylvania State University

We study the regression relationship among covariates in case-control data, an area known as the secondary analysis of case-control studies. The context is such that only the form of the regression mean is specified, so that we allow an arbitrary regression error distribution, which can depend on the covariates and thus can be heteroscedastic. Under mild regularity conditions we establish the theoretical identifiability of such models. Previous work in this context has either (a) specified a fully parametric distribution for the regression errors, (b) specified a homoscedastic distribution for the regression errors, (c) has specified the rate of disease in the population (we refer this as true population), or (d) has made a rare disease approximation. We construct a class of semiparametric estimation procedures that rely on none of these. The estimators differ from the usual semiparametric ones in that they draw conclusions about the true population, while technically operating in a hypothetic superpopulation.

email: yanyuanma@yahoo.com

113. BAYESIAN METHODS FOR SPATIAL AND SPATIO-TEMPORAL MODELING OF HEALTH DATA

STEP CHANGE DETECTION AND FORECASTING OF VECTOR-BORNE DISEASES

Gavino Puggioni*, University of Rhode Island
Jing Wu, University of Rhode Island

Recorded cases of tropical vector-borne diseases, such as Dengue and Zika, have been increasing in the last ten years and linked to climatic and anthropogenic changes. We propose new methods that address some common issues in these data: dynamic step change detection in spatial autocorrelation, and short and medium term forecast stability. The goal is to provide a flexible framework to identify and target areas with increased risk, and to inform early warning systems for surveillance and outbreak detection. We present a two stage space-time CAR model with Bayesian model averaging on the set of predictors, we apply it to a real dataset, and compare its forecasting performance with other widely used methods.

email: puggioni@cs.uri.edu

BAYESIAN DISAGGREGATION OF SPATIO-TEMPORAL COMMUNITY INDICATORS ESTIMATED VIA SURVEYS: AN APPLICATION TO THE AMERICAN COMMUNITY SURVEY

Veronica J. Berrocal*, University of Michigan
Marco H. Benedetti, University of Michigan

The American Community Survey (ACS) is an ongoing survey administered by the US Census Bureau which collects social, economic, and other community data. ACS estimates are released annually, with varying spatial and temporal resolution: 5-year estimates refer to smaller municipal subdivisions, while 1-year estimates refer to larger areas. Although for epidemiological studies, these estimates contain important community information, their varying spatial and temporal resolution pose various challenges: the 5-year ACS estimates might be temporally misaligned with finely resolved health outcome data, conversely, the coarser 1-year estimates are likely spatially misaligned with finely resolved health data. In this paper, we present a Bayesian hierarchical model that leverages both 1-year and 5-year ACS data and accounts for the survey sampling design to obtain estimates of community indicators at any given spatial and temporal resolution. The disaggregation is achieved by introducing a latent, point-referenced process, modeled using a multi-resolution basis function expansion and linked to the ACS data via a stochastic model that accounts also for the sampling survey design.

email: berrocal@umich.edu

AGE-SPECIFIC DISTRIBUTED LAG MODELS FOR HEAT-RELATED MORTALITY

Matthew J. Heaton*, Brigham Young University
Cassandra Olenick, National Center for Atmospheric Research
Olga V. Wilhelmi, National Center for Atmospheric Research

Distributed lag models have been consistently used throughout the years to assess the cumulative impact of multiple days of high heat on public health. Distributed

lag models, however, are often used on an aggregate level (e.g. city- or census-tracts) in spite of the fact that the effect of heat on health is individual-specific leading to possible ecological fallacies. To capture more individualized effects of heat on health, we propose a negative binomial regression model where the effect of lagged temperatures is age-specific. Utilizing principles of predictive process models, we place appropriate smoothness and decay constraints on the associated distributed lag surfaces. Further, we borrow strength across ages to facilitate the estimation of the age-specific distributed lag function. Using data from the Houston department of public health, we show how different ages are impacted by consecutive days of high heat.

email: mheaton@stat.byu.edu

RESTRICTED NONPARAMETRIC MIXTURES MODELS FOR DISEASE CLUSTERING

Abel Rodriguez*, University of California, Santa Cruz
Claudia Wehrhahn, University of California, Santa Cruz

Identifying disease clusters (areas with an unusually high incidence of a particular disease) is a common problem in epidemiology and public health. We describe a Bayesian nonparametric mixture model for disease clustering that constrains clusters to be made of contiguous areal units. This is achieved by modifying the exchangeable partition probability function associated with the Ewen's sampling distribution. The model is illustrated using data on cancer rates.

email: abel@soe.ucsc.edu

114. RECENT ADVANCES IN CAUSAL INFERENCE FOR SURVIVAL ANALYSIS

ADJUSTING FOR TIME-VARYING CONFOUNDERS IN SURVIVAL ANALYSIS USING STRUCTURAL NESTED CUMULATIVE SURVIVAL TIME MODELS

Stijn Vansteelandt*, Ghent University and London School of Hygiene and Tropical Medicine
Shaun Seaman, Cambridge University
Oliver Dukes, Ghent University
Ruth Keogh, London School of Hygiene and Tropical Medicine

Assessing the effect of a time-varying exposure on a survival endpoint is challenging. Standard survival methods that incorporate time-varying confounders as covariates generally yield biased estimates. Methods using inverse probability weighting are prone to giving unstable estimates when confounders are highly predictive of exposure or the exposure variable is continuous. Structural nested accelerated failure time models require artificial censoring, which can lead to estimation difficulties. Here, we introduce the structural nested cumulative survival time model (SNCSTM). This model assumes that setting exposure at time t to zero has an additive effect on the conditional hazard given the exposure and confounder histories when all subsequent exposures have already been set to zero. We show how SNCSTMs can be fitted using standard software and also describe two more efficient, double robust estimators, both available in closed form. All three estimators avoid artificial censoring and the instability of inverse weighting estimators. We examine the performance of our estimators using a simulation study and illustrate their use on data from the UK Cystic Fibrosis Registry.

email: stijn.vansteelandt@ugent.be

INSTRUMENTAL VARIABLES ESTIMATION WITH COMPETING RISK DATA

Torben Martinussen*, University of Copenhagen
Stijn Vansteelandt, Ghent University

Time-to-event analyses are often plagued by both – possibly unmeasured – confounding and competing risks. To deal with the former, the use of instrumental variables for effect estimation is rapidly gaining ground. We show how to make use of such variables in competing risk analyses. In particular, we show how to infer the effect of an arbitrary exposure on cause-specific hazard functions under a semi-parametric model that imposes relatively weak restrictions on the observed data distribution. The proposed approach is flexible accommodating exposures and instrumental variables of arbitrary type, and enabling covariate adjustment. It makes use of closed-form estimators that can be recursively calculated, and is shown to perform well in simulation studies. We also demonstrate its use in an application on the effect of mammography screening on the risk of dying from breast cancer. At the end of the talk I will discuss whether the popular 2SLS approach can be used in the setting of competing risk data.

email: tma@sund.ku.dk

INSTRUMENTAL VARIABLE ESTIMATION OF A COX MARGINAL STRUCTURAL MODEL WITH ENDOGENOUS TIME-VARYING EXPOSURE

Yifan Cui*, The Wharton School, University of Pennsylvania
Haben Michael, The Wharton School, University of Pennsylvania
Eric Tchetgen Tchetgen, The Wharton School, University of Pennsylvania

Robins (1998) introduced marginal structural models (MSMs), a general class of counterfactual models for the joint effects of time-varying treatment regimes in complex longitudinal studies subject to time-varying confounding. He established the identification of MSM parameters under a sequential randomization assumption (SRA), which rules out unmeasured confounding of treatment assignment over time. The Cox marginal structural model, in particular, is one of the most popular MSMs for evaluating the causal effect of a binary exposure with a censored failure time outcome. In this paper, we consider sufficient conditions for identification of Cox MSM parameters with the aid of a time-varying instrumental variable, when sequential randomization fails to hold due to unmeasured confounding. Our identification conditions essentially require that no interactions between unmeasured confoundings and the instrumental variable in its additive effects on the treatment, the longitudinal generalization of the identifying condition of Wang et al. (2018). Our approach is illustrated via simulation studies and a data analysis.

email: cuiy@wharton.upenn.edu

115. NOVEL STATISTICAL METHODS FOR ANALYSIS OF MICROBIOME DATA

HIGH DIMENSIONAL MEDIATION MODEL FOR MICROBIAL ABUNDANCE DATA

Ni Zhao*, Johns Hopkins University
Junxian Chen, The Hong Kong Polytechnic University

Emerging evidence has suggested the role of gut microbiome in the etiology of many health conditions, such as obesity, diabetes and hypertension. Recently, there has been increasing research interest in mediation analysis of microbiome data, i.e., assessing whether and to what extend the effect of an exposure (such as diet, medication) on a phenotype is transmitted through perturbing the microbial communities. Understanding this will provide us additional insight on the disease etiology, and potentially, lead to novel interventions for disease prevention and treatment. However, mediation analysis for microbiome data is challenging, due to the high dimensionality, sparsity and the complex interaction structure of the data. In this paper, we developed a high dimensional mediation model for microbial abundance data. Extensive simulations show that our method can lead to unbiased estimate of the mediation effect, and control type I type well. We further applied our approach to an intervention trial to investigate the effect of microbiome in mediating the effect of high fiber diet on reducing blood pressure.

email: nzhao10@jhu.edu

A SPARSE CAUSAL MEDIATION MODEL FOR MICROBIOME DATA ANALYSIS

Huilin Li*, New York University
Chan Wang, New York University
Jiyuan Hu, New York University
Martin Blaser, New York University

An important hallmark of human microbiota is its modifiability and dynamics. Many microbiome association studies have revealed the important association between microbiome and disease/health status, which encourage people to dive deeper to uncover the causation of microbiota in the underlying biological mechanism. Here, we propose a rigorous Sparse Causal Mediation Model specifically designed for the high dimensional and compositional microbiome data in a typical three-factor (treatment, microbiome and outcome) causal study design. In particular, the Dirichlet multivariate regression model and linear log-contrast regression model are proposed to estimate the causal direct effect of treatment and mediation effect of the microbiota at both community level and individual level. Regularization techniques are used to perform the variable selection in the proposed model framework to identify signature causal microbes. Simulations and real data analyses are used to evaluate the performance of the proposed method.

email: huilin.li@nyumc.org

A ROBUST AND POWERFUL FRAMEWORK FOR MICROBIOME BIOMARKER DISCOVERY

Jun Chen*, Mayo Clinic
Li Chen, Auburn University

One central theme of microbiome studies is to identify bacterial taxa/functions associated with some clinical or biological outcome (a.k.a, microbiome biomarker discovery). Many methods have been proposed for this task, ranging from simple Wilcoxon rank sum test to sophisticated zero-inflated models. Due to the excessive zeros, outliers, compositionality and phylogenetic structure in microbiome data, existing methods are still far from optimal: parametric methods tend to be less robust while non-parametric methods are less powerful. We thus propose a permutation test for robust and powerful microbiome biomarker discovery. The method is based on weighted least squares estimation for linear models and hence is computationally efficient. It takes into account the full characteristics of microbiome sequencing data including variable library sizes, the correlations among taxa, the inherent compositionality, and the phylogenetic relatedness of the taxa. An omnibus test is designed to capture various biological effects. We demonstrate the robustness and power of our method by extensive simulations and real data applications.

email: chen.jun2@mayo.edu

OMNIDIRECTIONAL VISUALIZATION OF COMPETITION AND COOPERATION IN THE GUT MICROBIOTA

Rongling Wu*, The Pennsylvania State University

Microbial interactions in the gut are thought to determine host health, but the visualization of how interacting architecture forms in the gut microbiota is unknown. Here, we developed an approach for constructing bidirectional, signed and weighted networks for microbial community assembly. We found that the abundance of one microbe (part) scales allometrically with the total abundance of all microbes (whole) across ecological niches. We married this common power law and evolutionary game theory to characterize how individual microbes interact dynamically with each other by choosing optimal competition or cooperation strategies. Variable selection models were implemented to select the most significant subset of microbes with which a focal microbe is linked, ultimately leading to a sparse but omnidirectional network of microbial interactions. The approach was applied to analyze a published gut microbiota data collected from an isolated human population, enabling the global visualization of internal workings in gut microbiomes.

email: rwu@phs.psu.edu

116. NEW DEVELOPMENTS IN NONPARAMETRIC METHODS FOR COVARIATE SELECTION

VARIABLE PRIORITIZATION IN BLACK BOX STATISTICAL METHODS

Lorin Crawford*, Brown University

The central aim of this talk is to address variable selection questions in nonlinear and nonparametric regression. Motivated by statistical genetics, where nonlinear

interactions are of particular interest, we introduce a novel and interpretable way to summarize the relative importance of predictor variables. Methodologically, we develop the "RelATive cEntrality" (RATE) measure to prioritize candidate genetic variants that are not just marginally important, but whose associations also stem from significant covarying relationships with other markers in the data. We illustrate RATE through Bayesian Gaussian process regression, but the methodological innovations apply to other "black box" methods (e.g. deep neural networks). It is known that nonlinear models often exhibit greater predictive accuracy than linear models, particularly for phenotypes generated by complex genetic architectures. With detailed simulations and real data association mapping studies, we show that applying RATE enables an explanation for this improved performance.

email: lorin_crawford@brown.edu

COVARIATE SELECTION AND ALGORITHMIC FAIRNESS FOR CONTINUOUS OUTCOMES IN HEALTH PLAN RISK ADJUSTMENT

Sherri Rose*, Harvard Medical School
Anna Zink, Harvard University

Risk adjustment formulas aim to predict spending in health insurance markets in order to provide fair benefits and health care coverage for all enrollees, regardless of their health status. Unfortunately, current risk adjustment formulas are known to undercompensate payments to health insurers for specific groups of enrollees (by underpredicting their spending). In this talk, we expand concepts from both the computer science and health economics literature to develop definitions of fairness for continuous outcomes in health plan payment risk adjustment formulas. We additionally propose new estimation methods in an effort to make risk adjustment fairer for undercompensated groups. While adding risk adjustors is a common policy approach, it is often not effective. Our data application demonstrates that alternative methods improve risk adjustment formulas with respect to an undercompensated group, and that a suite of metrics is necessary in order to evaluate the formulas more fully.

email: rose@hcp.med.harvard.edu

NON-PARAMETRIC AND DATA-DRIVEN METHODS FOR IDENTIFYING SUBPOPULATIONS SUSCEPTIBLE TO THE HEALTH EFFECTS OF AIR POLLUTION

Cole Brokamp*, Cincinnati Children's Hospital Medical Center

Research utilizing large datasets including electronic health records (EHR) and other administrative databases have allowed for an unprecedented examination of population-level health risks associated with exposure to air pollution. While the population-level approach is useful with respect to policy, there may be individuals or subgroups of the population particularly susceptible to air pollution. The National Ambient Air Quality Standards (NAAQS) state that standards must be sufficiently protective for at-risk populations, but reproducible and data-driven statistical methods for identifying these subpopulations are critically missing. Here, we accomplish this by building on transformed outcome forests and extending them to deal with confounding in observational studies. Using a database of over 35 individual-level susceptibility characteristics and a population-wide pediatric EHR study population, we demonstrate the utility of our method by identifying subpopulations that are the most susceptible to acute air pollution exposures with respect to psychiatric emergency department encounters.

email: cole.brokamp@cchmc.org

DYNAMIC LANDMARK PREDICTION FOR MIXTURE DATA

Tanya P. Garcia*, Texas A&M University
Layla Parast, RAND Corporation

In kin-cohort studies, clinicians are interested in providing their patients with the most current cumulative risk of death arising from a rare deleterious mutation. Estimating the cumulative risk is difficult when the genetic mutation status in patients is unknown and only estimated probabilities of a patient having the mutation are available. We estimate the cumulative risk for this scenario using a new nonparametric estimator that incorporates covariate information and dynamic landmark prediction. Our estimator more precisely predicts the risk of death compared to existing estimators that ignore covariate information. Our estimator is built within a dynamic landmark prediction framework whereby we obtain personalized dynamic predictions over time. A simple transformation of our estimator also provides more efficient estimates of marginal distribution functions in settings where patient-specific predictions are not necessarily the main goal. Applying our method to a Huntington disease study, we develop survival prediction curves incorporating gender and familial genetic information, and create personalized dynamic risk trajectories over time.

email: tpgarcia@stat.tamu.edu

117. DYNAMIC TREATMENT REGIMENS AND EXPERIMENTAL DESIGN

SHOULD I STAY OR SHOULD I GO: SELECTING INDIVIDUALIZED STAGE DURATION IN A SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIAL (SMART)

Hayley M. Belli*, New York University Langone School of Medicine
Andrea B. Troxel, New York University Langone School of Medicine

In a Sequential Multiple Assignment Randomized Trial (SMART), each participant is randomly assigned a treatment, but following this initial assignment, the patient moves through series of stages with the option to continue or switch interventions based on response. With time, patients will be assigned to more effective treatments. The length of the initial treatment period, however, is fixed and usually selected by investigators in advance without much guiding data or information; this timing is also applied uniformly to all participants in the study. For example, if a stage lasts three months, a patient who has early information about the lack of effectiveness of a treatment, say at one month, would still receive an ineffective treatment unnecessarily for two months. The present work addresses this shortcoming of SMARTs by introducing regression and likelihood-based methods to determine when an individual patient should stay or switch interventions. As we strive to advance precision medicine, this approach uses data and design principles to inform the duration of stages to maximize trial benefit and provide personalized patient care all within a rigorous experimental framework.

email: hayley.belli@nyulangone.org

NEW STATISTICAL LEARNING FOR EVALUATING NESTED DYNAMIC TREATMENT REGIMES

Ming Tang*, University of Michigan
Lu Wang, University of Michigan
Jeremy M.G. Taylor, University of Michigan

Dynamic treatment regimes (DTRs) are sequences of treatment decision rules based on the changing of one's disease progression and healthcare history. In medical practice, nested treatment/exam(s) assignments are common to improve cost-effectiveness. For example, people at risk of prostate cancer have their Prostate-specific antigen (PSA) tested regularly. Only ones have high index in PSA need biopsy exam, which is costly and invasive, to confirm the treatment assignment. Treatment switch happens after biopsy, and is thus nested within the decision of taking biopsy. We develop statistical method to evaluate DTR within such a nested multi-stage dynamic decision framework. At each step of each stage, we combine semi-parametric estimation via Augmented Inverse Probability Weighting with tree-based reinforcement learning to deal with the counterfactual optimization. The simulation study shows the robust performances of the proposed methods under different scenarios. We further apply our method in evaluating the necessity of prostate biopsy and identifying the optimal treatment following biopsy positive results for prostate cancer patients in the active surveillance system.

email: mingtang@umich.edu

DESIGN AND ANALYSIS ISSUES FOR ESTIMATING TRANSMISSION PROBABILITIES IN A CHALLENGE STUDY

Sally Hunsberger*, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Michael A. Proschan, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Alison Han, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Matthew J. Memoli, National Institute of Allergy and Infectious Diseases, National Institutes of Health

Healthy volunteer influenza challenge studies, where cohorts of participants are given a wild-type influenza virus, have been used to study pathogenesis of influenza and evaluate efficacy in phase II trials of new vaccines and therapeutics. Recently, interest in studying transmission in these studies has arisen, as studies can be designed where some members of a cohort are challenged with virus, while others receive placebo. The goals of such a study are to demonstrate that the virus can be transmitted, identify methods of transmission, and to estimate the probability of transmission. Successful design of these studies requires careful consideration of a number of factors including estimation of transmission probabilities. Here we investigate numerical design issues such as the optimal number of patients in a cohort to challenge in order to have a high probability of observing at least one transmission event and the number of cohorts necessary. We also consider statistical aspects such as the mean-squared error (MSE) of the estimated transmission probability and the construction of confidence intervals for the transmission probability.

email: sallyh@mail.nih.gov

DISCOVERY OF GENE REGULATORY NETWORKS USING ADAPTIVELY-SELECTED GENE PERTURBATION EXPERIMENTS

Michele S. Zemplenyi*, Harvard University
Jeffrey W. Miller, Harvard University

Graphical models have previously been used to reconstruct gene networks from RNA sequencing data. However, since several graphs can often explain data equally well, not all causal relationships can be inferred from observational data alone. Instead, perturbation experiments, such as gene knock-outs, are needed. Because different perturbation experiments yield varying degrees of information about the causal structure of a network, it is advantageous to select perturbations that most efficiently narrow down the set of possible causal graphs. In particular, we wish to find the optimal sequence of experiments that will yield the greatest gain of information about a gene network. To this end, we use MCMC methods to explore the space of networks and to estimate the reduction in entropy that would result from a particular perturbation. This ability to adaptively select experiments, combined with recent advances in the precision of gene-knockout experiments, provides a promising avenue for reconstructing gene networks by iterating between experimentation and analysis. We compare our learning algorithm to alternative perturbation selection schemes via a simulation study.

email: mzemplenyi@g.harvard.edu

A SAMPLE SIZE CALCULATION FOR BAYESIAN ANALYSIS OF SMALL N SEQUENTIAL MULTIPLE ASSIGNMENT RANDOMIZED TRIALS (snSMARTs)

Boxian Wei*, University of Michigan
Thomas M. Braun, University of Michigan
Roy N. Tamura, University of South Florida
Kelley M. Kidwell, University of Michigan

A clinical trial design for rare diseases is the small n Sequential Multiple Assignment Randomized Trial (snSMART), in which subjects are first randomized to one of the multiple treatments (stage1). Responders continue the same treatment for another stage, while non-responders are re-randomized to one of the remaining treatments (stage2). A Bayesian analysis for snSMARTs shows efficiency gains in treatment efficacy estimation. As limited sample size calculations for snSMARTs exist, we propose a Bayesian sample size calculation for an snSMART designed to distinguish the best from the second-best treatment. Although our methods are based on asymptotic approximations, we show via simulation that our proposed sample size calculation produces the desired statistical power, even in small samples. We also compare our proposed sample size to those produced from (a) a two-stage method using weighted Z-statistics and (b) a standard sample size calculation for a single stage parallel group clinical trial. Given the same desired coverage rate/type I error and statistical power, we show that the Bayesian analysis requires fewer subjects than competing approaches.

email: boxian@umich.edu

DESIGN OF EXPERIMENTS FOR A CONFIRMATORY TRIAL OF PRECISION MEDICINE

Kim May Lee*, University of Cambridge
James Wason, University of Cambridge

Precision medicine is becoming more pronounced in the medical field as genomic information could be used to refine cancer treatments such that different treatments may provide maximum benefit to the heterogeneous patients. These heterogenous patients could be stratified into several subgroups based on a measured indicator known as a biomarker. A multi-arm trial could be implemented if it is known that which subgroups would be benefited from the treatment regimes, otherwise enrolling all patient subgroups in a confirmatory trial would increase the burden of the study. We propose a design framework for finding an optimal design that could be implemented in a phase III study or a confirmatory trial based on the results of a phase II study. We use Bayesian data analysis to select subgroups and treatments to be enrolled in the future trial, and experimental design framework to find the optimal treatment randomization scheme. We study the characteristics of the designs using simulation and explore the efficiency loss of using other standard designs when instead the optimal design should have been used.

email: kim.lee@mrc-bsu.cam.ac.uk

118. HYPOTHESIS TESTING AND SAMPLE SIZE CALCULATION

BAYESIAN NONPARAMETRIC TEST FOR INDEPENDENCE BETWEEN RANDOM VECTORS

Zichen Ma*, University of South Carolina
Timothy E. Hanson, University of South Carolina

Hypothesis testing for independence between groups of random variables, especially when the dependence is nonlinear, is both of theoretical interest and crucial in applications. We propose a nonparametric approach to independence testing between sets of continuous random variables. Multivariate finite Polya tree priors are used to model the underlying probability distributions. Integrating out the random probability measure, we derive a tractable empirical Bayes factor as the test statistic and the basis for obtaining a p-value through a standard permutation test. Further, we develop a measure of dependence based on the total variation norm that quantifies how related the sets of random variables are, serving a similar purpose as the classic dependence measures such as Pearson's correlation. To demonstrate the ability of the proposed method, we provide a set of simulation studies under both linear and nonlinear dependency in which our approach is compared to several existing approaches. Lastly, we provide a real data application of the proposed method to a data set from ecology where we test whether the spread of a certain fungus exhibits spatial-temporal dependency.

email: zichen@email.sc.edu

KERNEL BASED-HYBRID TEST FOR HIGH-DIMENSIONAL DATA

Inyoung Kim*, Virginia Tech University

Numerous statistical methods have been developed for analyzing high-dimensional data. These methods often focus on variable selection approaches but are limited for the purpose of testing with high-dimensional data. They are often required to have explicit likelihood functions. In this paper, we propose a kernel based-hybrid omnibus test for high-dimensional data testing purpose with much weaker requirements. Our hybrid omnibus test is developed under a semiparametric framework where a likelihood function is no longer necessary. Our test is a version of a frequentist-Bayesian hybrid score-type test for a nonparametric model, which has a link function being a functional of a set of variables through Gaussian processes with high correlated variables. We propose an efficient score based on estimating equations and then construct our hybrid omnibus test using local tests. We compare our approach with an empirical likelihood ratio test and Bayesian inference based on Bayes factors, using simulation studies. The advantage of our approach is demonstrated by applying it to genetic pathway data for type II diabetes mellitus.

email: inyoungk@vt.edu

A NON-NESTED HYPOTHESIS TESTING PROBLEM FOR THRESHOLD REGRESSION MODELS

Zonglin He*, Fred Hutchinson Cancer Research Center

Motivated by an example from the prevention of mother-to-child transmission of HIV-1, we study a non-nested hypothesis testing problem that seeks to discriminate between a linear model and a hinge model, which is a special type of threshold model. To develop the test statistic, we follow the same approach as used in the nested testing problem; to obtain p-values, the approach is necessarily different. We study three parametric bootstrap procedures and find that fast double parametric bootstrap procedures have close to nominal type 1 error rates and are reasonably fast. Both linear regression and logistic regression are considered. We illustrate the proposed methods using a dataset from the motivating application.

email: nkhezl@gmail.com

ROBUST BOOTSTRAP TESTING FOR NONLINEAR EFFECT IN SMALL SAMPLE WITH KERNEL ENSEMBLE

Wenying Deng*, Harvard T.H. Chan School of Public Health
Jeremiah Zhe Liu, Harvard T.H. Chan School of Public Health
Brent Coull, Harvard T.H. Chan School of Public Health

Kernel machine-based hypothesis tests have seen widespread application in GWAS and gene-environment interaction studies. However, testing for high-dimensional, nonlinear interaction effect between groups of continuous features remains difficult in practice. The main challenge roots from constructing an efficient and unbiased estimator for the complex, nonlinear main-effect model. The recently proposed Cross-Validated Ensemble of Kernels (CVEK) addresses this challenge by proposing an ensemble-based estimator that adaptively learns the form of the main-effect kernel from data, and constructs an companion variance component test. However, the null distribution of CVEK relies on asymptotic approximation, therefore calling

into question the validity of the test in limited sample. In this work, we address this shortcoming by proposing a bootstrap test for CVEK. We conduct comprehensive simulation to compare the validity (i.e. Type I error) and power of both bootstrap and asymptotic test, and reveal novel insight on the performance of variance component test under nonlinear data-generation mechanisms.

email: wdeng@hsph.harvard.edu

A ROBUST HYPOTHESIS TEST FOR CONTINUOUS NONLINEAR INTERACTIONS IN NUTRITION-ENVIRONMENT STUDIES: A CROSS-VALIDATED ENSEMBLE APPROACH

Jeremiah Zhe Liu*, Harvard T.H. Chan School of Public Health
Jane Lee, Boston Children's Hospital
Pi-i Debby Lin, Harvard University
Linda Valeri, Columbia Mailman School of Public Health
David Christiani, Harvard T.H. Chan School of Public Health
David Bellinger, Harvard T.H. Chan School of Public Health
Robert Wright, Icahn School of Medicine at Mount Sinai
Maitreyi Mazumdar, Boston Children's Hospital
Brent Coull, Harvard T.H. Chan School of Public Health

Gene-environment and nutrition-environment studies often involve testing of high-dimensional interactions between two sets of variables, each having potentially complex nonlinear main effects on an outcome. Construction of a valid and powerful hypothesis test for such an interaction is challenging, due to the difficulty in constructing an efficient and unbiased estimator for the complex, nonlinear main effects. In this work we address this problem by proposing a Cross-validated Ensemble of Kernels (CVEK) that learns the space of appropriate functions for the main effects using a cross-validated ensemble approach. With a carefully chosen library of base kernels, CVEK flexibly estimates the form of the main-effect functions from the data, and encourages test power by guarding against over-fitting under the alternative. The method is motivated by a study on the interaction between metal exposures in utero and maternal nutrition on children's neurodevelopment in rural Bangladesh. The proposed tests identified evidence of an interaction between minerals and vitamins and Arsenic and Manganese exposures, and was more powerful than existing approaches.

email: zhl112@mail.harvard.edu

A GOODNESS OF FIT TEST TO COMPARE LUMPED AND UNLUMPED MARKOV CHAINS

Anastasia M. Hartzes*, University of Alabama at Birmingham
Charity J. Morgan, University of Alabama at Birmingham

It is common practice to aggregate categorical outcomes for statistical or sample size considerations; in the case of Markov Chains (MC), combining states is referred to as lumping. After lumping, chains must be evaluated to ensure Markov dependency is maintained, for which a Chi-square test of lumpability exists. As an extension, we propose a goodness of fit (GOF) test which compares lumped and unlumped chains. While a simpler matrix could be preferable, a more appropriate and larger one might provide greater insight and superior fit to the data. Using semi-annual surveys from 2007-12 from the North American Research Committee on Multiple Sclerosis (NARCOMS), marital status was modeled via MC. Transition matrices were simulated to assess the GOF statistic, which was shown to follow a Chi-square distribution. While performed for a simple comparison of transition matrices, the method would ideally be applied to more complex models that account for covariates. Lumping states can ease interpretability and computation. However, it must be justified by scientific and statistical considerations, including confirmation that chain is still Markov.

email: ahartzes@uab.edu

SAMPLE SIZE FOR TRIALS COMPARING GROUP AND INDIVIDUAL TREATMENTS WITH REPEATED MEASURES

Robert J. Gallop*, West Chester University

Partially clustered designs, where clustering occurs in some conditions and not others, are common in psychology, particularly in area of prevention and intervention trials. Baldwin, Stice, & Rohde (2008) discussed the analysis of these designs and Moerbeek and Wong (2008) produced power formulas for these partially clustered designs, which are referred to as 2-1 models, with one treatment arm has two levels (patients within group) and the other is one level (patients alone). In this presentation, we extend power formula to consist of an additional level, repeated observations per patient. We refer to this structure as a 3-2 design. The power derivations consists of two-stages: the first for repeated measures and the second for the 3-2 mixed-group design. Under two extremes: either no repeated measures or within-subject correlation of 1, we show the proposed power formulas reduce to the 2-1 formulas. We will illustrate the power derivation as well as sensitivity of the calculations illustrated for two studies: Cocaine addiction study and a Depression prevention intervention for adolescents.

email: rgallop@wcupa.edu

119. MEASUREMENT ERROR

CATEGORIZING A CONTINUOUS PREDICTOR SUBJECT TO MEASUREMENT ERROR

Tianying Wang*, Columbia University
Raymond Carroll, Texas A&M University
Betsabe Blas Achic, Universidade Federal de Pernambuco
Ya Su, University of Kentucky
Victor Kipnis, National Cancer Institute, National Institutes of Health
Kevin Dodd, National Cancer Institute, National Institutes of Health

Epidemiologists often categorize a continuous risk predictor, even when the true risk model is not a categorical one. Nonetheless, such categorization is thought to be more robust and interpretable, and thus their goal is to fit the categorical model and interpret the categorical parameters. We address the question: with measurement error and categorization, how can we do what epidemiologists want, namely to estimate the parameters of the categorical model that would have been estimated if the true predictor was observed? We develop a general methodology for such an analysis, and illustrate it in linear and logistic regression. Simulation studies are presented and the methodology is applied to a nutrition data set. Discussion of alternative approaches is also included.

email: tw2696@cumc.columbia.edu

EFFICIENT INFERENCE FOR TWO-PHASE DESIGNS WITH RESPONSE AND COVARIATE MEASUREMENT ERROR

Sarah C. Lotspeich*, Vanderbilt University
Bryan E. Shepherd, Vanderbilt University Medical Center
Pamela Shaw, University of Pennsylvania
Ran Tao, Vanderbilt University Medical Center

In modern electronic health records systems, both the outcome and covariates of interest can be error-prone, and these errors are often correlated. A cost-effective solution is the two-phase design, under which the error-prone outcome and covariates are observed for all subjects during the first phase and that information is used to select a validation subsample for accurate measurements of these variables in the second phase. Previous research on two-phase measurement error problems largely focused on scenarios where there are errors in covariates only or the validation sample is a simple random sample. Herein, we propose a semiparametric approach to general two-phase measurement error problems with a continuous outcome, allowing for correlated errors in the outcome and covariates and arbitrary second-phase selection. We devise a computationally efficient and numerically stable EM algorithm to maximize the nonparametric likelihood function. The resulting estimators possess desired statistical properties. We compare the proposed methods to existing approaches through extensive simulation studies, and we illustrate their use in an observational HIV study.

email: sarah.c.lotspeich@vanderbilt.edu

IMPROVING THE REPRODUCIBILITY OF EHR-BASED ASSOCIATION STUDIES FOR PLEIOTROPIC EFFECTS BY ACCOUNTING FOR PHENOTYPING ERRORS

Jiayi Tong*, University of Pennsylvania
Ruowang Li, University of Pennsylvania
Doudou Zhou, University of Science and Technology of China
Rui Duan, University of Pennsylvania
Jason Moore, University of Pennsylvania
Yong Chen, University of Pennsylvania

Pleiotropy, which occurs when one genetic locus influences multiple phenotypes, has the potential to uncover the complex mechanisms between genotypes and phenotypes. With the increasingly more electronic health record (EHR) systems starting to genotype patients in their linked biobanks, it is now feasible to assess pleiotropic effects in many clinical phenotypes with adequate sample sizes. Despite the overwhelmingly amount of phenotypes in EHR, the quality of those phenotypes are mired by various types of bias and errors. In particular, the phenotyping errors in EHR data can greatly negatively impact the reproducibility of EHR-based findings including identification of pleiotropic effects. In the paper, we propose a novel procedure for handling the potentially error-prone phenotypes (i.e. the surrogate) obtained from high-throughput phenotyping algorithms to estimate pleiotropic effects with EHR-based data. Using Penn Medicine Biobank (PMBB) EHR data, we demonstrate that the proposed method can lead to much improved statistical power while maintaining well controlled type I errors in identifying true pleiotropic effects compared to existing methods.

email: ychen123@mail.med.upenn.edu

BAYESIAN LATENT CLASS REGRESSION FOR MEASUREMENT ERROR CORRECTION IN SELF-REPORTED DIETARY INTAKE

Caroline P. Groth*, Northwestern University Feinberg School of Medicine
David Aaby, Northwestern University Feinberg School of Medicine
Michael J. Daniels, University of Florida
Linda Van Horn, Northwestern University Feinberg School of Medicine
Juned Siddique, Northwestern University Feinberg School of Medicine

Measurement error correction of self-reported diet requires information on the relationship between self-reported diet and its true value. This information is usually assessed by a validation study that measures 24-hour urinary biomarkers and self-reported diet which is then applied to a new study that lacks biomarkers. Assuming a single relationship between a variable measured with error and its true value may undermine correction of self-reported diet due to subject variation. To better account for different measurement error relationships, we developed a Bayesian latent class regression model that forms classes that represent different measurement error relationships between biomarkers and self-reported diet. We allow class membership to be determined by characteristics that influence measurement error. Our model is then applied to studies without biomarkers to correct for measurement error and assess the relationship between intake and health outcomes. We display our method by fitting a measurement error correction model for self-reported sodium intake using data from four validation studies which collected biomarkers and applying it to self-reported sodium in INTERMAP.

email: caregroth@gmail.com

BAYESIAN APPROACH FOR HANDLING COVARIATE MEASUREMENT ERROR WHEN ESTIMATING POPULATION TREATMENT EFFECT

Hwanhee Hong*, Duke University
Juned Siddique, Northwestern University Feinberg School of Medicine
Elizabeth A. Stuart, Johns Hopkins Bloomberg School of Public Health

Randomized controlled trials (RCTs) are the gold standard for evaluating intervention effects. However, results of RCTs may not be generalizable to a target population for which we want to make decisions regarding treatment implementation. Measurement error can be easily found in either RCT or target population data, but methods for handling it in the generalizability context have not been developed. In this talk, we propose a flexible Bayesian approach for handling such covariate measurement error when estimating population treatment effects. Bayesian hierarchical models impute the unobserved true covariate by learning the measurement error structure from the validation data. We assess the performance of our methods via simulations. We consider scenarios where an error-prone covariate exists in (1) both RCT and target population, (2) RCT only, and (3) target population only. We investigate these scenarios with and without validation data in either trial or population data. We apply our methods to a real data example to assess the population treatment effect of a program to reduce sodium intake on hypertension using PREMIER (RCT) and INTERMAP studies (target population).

email: hwanhee.hong@duke.edu

FLEXIBLE OMNIBUS TEST IN 1:M MATCHED CASE-CROSSOVER STUDY WITH MEASUREMENT ERROR IN COVARIATE

Byung-Jun Kim*, Virginia Tech University
Inyoung Kim, Virginia Tech University

The matched case-crossover study is a popular design in epidemiological research with clustered binary outcomes. In such studies, some matching covariates often play an important role as effect modifications causing incorrect statistical testing. covariates are often measured with error as well. Not accounting for this error can also lead to incorrect inferences for all covariates in the model. The methods for evaluating effect modification by matching covariates as well as assessing error-in-covariates in matched case-control studies are quite limited. Thus, we propose a flexible omnibus test which can be used for testing (1) the significance of a functional association between the clustered binary outcomes and covariates with measurement error, (2) the existence of the effect modification by matching covariates, and (3) the significance of an interaction effect between measurement error covariate and other covariate. The proposed omnibus test has the flexibility to make inferences on various hypothesis settings. The advantages of our omnibus test are demonstrated through simulation studies and bidirectional matched data analyses from an epidemiology study.

email: bjkim702@vt.edu

120. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

INTEGRAL PROJECTION MODELS FOR POPULATION IN COLUMBIAN GROUND SQUIRREL

Kyoung Ju Kim*, Auburn University

Plant and animal could have substantial interannual variability in survival, growth, and fecundity. Understanding the connection between historical variation in climate and population vital rates is one way to predict the impact of future climate change. Using a data set for Columbian ground squirrel, we parameterized stochastic integral projection models (IPMs) to identify vital rates and climate variables, which are most important for population growth.

email: kjk0019@auburn.edu

A BAYESIAN CRITICAL WINDOW VARIABLE SELECTION METHOD FOR ESTIMATING THE IMPACT OF AIR POLLUTION EXPOSURE DURING PREGNANCY

Joshua L. Warren*, Yale University
Wenjing Kong, Yale University
Thomas J. Luben, United States Environmental Protection Agency
Howard H. Chang, Emory University

Understanding the impact that environmental exposure during different stages of pregnancy has on the risk of adverse birth outcomes is vital for the development of mechanistic explanations of exposure-disease relationships. Statistical models that seek to estimate critical windows of susceptibility have been developed for a number of different reproductive outcomes and pollutants. However, these current methods fail to adequately address the primary objective of interest; how to define a critical window of susceptibility. We introduce critical window variable selection (CWVS), a hierarchical Bayesian framework that directly addresses this question while simultaneously providing improved estimation of the risk parameters of interest. Through simulation, we show that CWVS outperforms existing competing techniques in the setting of highly temporally correlated exposures in terms of correctly identifying critical windows and accurately estimating the risk parameters. We apply all competing methods to a case/control analysis of pregnant women in North Carolina, 2005-2008, with respect to the development of very preterm birth and exposure to ambient ozone and PM2.5.

email: joshua.warren@yale.edu

COMBINING AIR POLLUTION ESTIMATES FROM MULTIPLE STATISTICAL MODELS USING SPATIAL BAYESIAN ENSEMBLE AVERAGING

Nancy L. Murray*, Emory University
Howard H. Chang, Emory University

Ambient fine particulate matter less than 2.5 μm in aerodynamic diameter (PM2.5) negatively affects respiratory health, cardiovascular health, and other health outcomes. The United States Environmental Protection Agency implements a national monitoring network for PM2.5; however, the spatial and temporal sparsity of the observed monitoring data often limits the scope of epidemiologic analyses. We develop a method to combine multiple current statistical data integration techniques that utilize PM2.5 monitoring data to predict daily concentrations at unmonitored locations. Via data augmentation techniques, we model spatially varying weights, informed by covariates, for each data integration technique. The resulting weights are then used in a Bayesian ensemble averaging framework to combine estimates across models, providing improved estimates compared to using individual data integration techniques. We apply the method to obtain estimates of PM2.5 at 1 km spatial resolution, along with estimates of uncertainty, in 2012 for the contiguous United States.

email: nancy.murray@emory.edu

A HIERARCHICAL MODEL FOR ESTIMATING EXPOSURE-RESPONSE CURVES FROM MULTIPLE STUDIES

Joshua P. Keller*, Colorado State University
Scott L. Zeger, Johns Hopkins University

Estimating exposure-response curves is critical for quantifying the impact of environmental exposures. In intervention trials targeting indoor air pollution, measurements of exposure and surveillance of outcomes are frequently clustered, sparse across time, and impacted by seasonal trends. Pooling data from multiple studies provides an opportunity to increase power for estimating the exposure-response relationship. We present a hierarchical approach to modeling exposure concentrations and combining data from multiple studies in order to estimate a common exposure-response curve while accounting for temporal heterogeneity and multiple levels of clustering. We apply this model to data from three studies of cookstoves and respiratory infections in children, which represent three study types: crossover trial, parallel trial, and case-control study.

email: joshua.keller@colostate.edu

ROBUST NONPARAMETRIC DERIVATIVE ESTIMATOR

Hamdy F. F. Mahmoud*, Virginia Tech University
Byung-Jun Kim, Virginia Tech University
Inyoung Kim, Virginia Tech University

In this paper, a robust nonparametric derivative estimator is proposed to estimate the derivative function of nonparametric regression when the data contain noise and have curves. A robust estimation of the derivative function is important for understanding trend analysis and conducting statistical inferences. The methods for simultaneously assessing the functional relationship between response and covariates as well as estimating its derivative function without trimming noisy data are quite limited. Our robust nonparametric derivative functions were developed by constructing three weights and then incorporating them into kernel-smoothing. Various simulation studies were conducted to evaluate the performance of our approach and to compare our proposed approach with other existing approaches. The advantage of our robust nonparametric approach is demonstrated using epidemiology data on mortality and temperature in Seoul, South Korea.

email: ehamdy@vt.edu

121. STATISTICAL METHODS FOR HIGH DIMENSIONAL DATA

GENERALIZED LINKED MATRIX FACTORIZATION

Michael J. O'Connell*, Miami University

Multi-source data, where multiple types of variables are collected for the same set of objects, have become common in many areas of research as data collection and storage has become easier with advancing technology. Linked matrices represent multi-source data sets with matrices that share dimensions. Matrices can be linked in two ways: sharing a common variable set (horizontal integration) or sharing a common sample set (vertical integration). Some data sets contain both horizontal and vertical integration. Linked Matrix Factorization (LMF) can be used for dimension reduction of such data. However, LMF assumes that all data sets are normally distributed. I will describe a generalized version of LMF for linked matrices that follow exponential family distributions, which is based on an alternating generalized linear models algorithm rather than an alternating least squares algorithm.

email: oconnemj@miamioh.edu

ESTIMATION AND INFERENCE FOR HIGH DIMENSIONAL GENERALIZED LINEAR MODELS: A SPLIT AND SMOOTHING APPROACH

Zhe Fei*, University of Michigan
Yi Li, University of Michigan

Understanding the molecular causes of the lung cancer heterogeneity is one of the main focuses in current cancer researches. The Boston Lung Cancer Study (BLCS) is a cohort of over 11,000 lung cancer cases with various genetic measurements. We propose a novel framework of estimation and inference for high dimensional generalized linear models that is adaptive to the large scale data structures in the BLCS. We use data splitting and smoothing to derive coefficient estimators with approximately normal asymptotics. The proposed procedure is significantly different from the majority of the related works in that penalization is not involved in the estimation to deal with the high dimensionality. Our method provides accurate point estimation as well as reliable inferences for any single coefficient or low dimensional subvectors of the coefficients. Extensive numerical experiments and the application to a BLCS SNP data are presented. We estimate and test both the SNP effects and their interactions with smoking, which leads to meaningful and novel findings.

email: feiz@umich.edu

ADAPTIVE SPARSE ESTIMATION WITH SIDE INFORMATION

Trambak Banerje*, University of Southern California
Gourab Mukherjee, University of Southern California
Wenguang Sun, University of Southern California

The article considers the problem of estimating a high-dimensional sparse parameter in the presence of auxiliary data that encode side information on sparsity. We develop a general framework that involves first constructing an auxiliary sequence to capture the side information, and then incorporating the auxiliary sequence in inference to reduce the estimation risk. The proposed method, which carries out adaptive SURE-thresholding using side information (ASUS), is shown to have robust performance and enjoy optimality properties. We develop new theories to characterize regimes in which ASUS far outperforms competitive shrinkage estimators, and establish precise conditions under which ASUS is asymptotically optimal. Simulation studies are conducted to show that ASUS substantially improves the performance of existing methods in many settings. The methodology is applied for analysis of data from single cell virology studies.

email: trambakb@usc.edu

COVARIATE ASSISTED PRINCIPAL REGRESSION FOR COVARIANCE MATRIX OUTCOMES

Yi Zhao*, Johns Hopkins Bloomberg School of Public Health
Bingkai Wang, Johns Hopkins Bloomberg School of Public Health
Stewart H. Mostofsky, Johns Hopkins University
Brian S. Caffo, Johns Hopkins Bloomberg School of Public Health
Xi Luo, Brown University

Modeling variances in data has been an important topic in many fields, including in financial and neuroimaging analysis. We consider the problem of regressing covariance matrices on a vector covariates, collected from each observational unit. The main aim is to uncover the variation in the covariance matrices across units that are explained by the covariates. This paper introduces Covariate Assisted Principal (CAP) regression, an optimization-based method for identifying the components predicted by (generalized) linear models of the covariates. We develop computationally efficient algorithms to jointly search the projection directions and regression coefficients, and we establish the asymptotic properties. Using extensive simulation studies, our method shows higher accuracy and robustness in coefficient estimation than competing methods. Applied to a resting-state functional magnetic

resonance imaging study, our approach identifies the human brain network changes associated with age and sex.

email: zhaoyi1026@gmail.com

ESTIMATING T-CENTRAL SUBSPACE VIA MARGINAL THIRD MOMENTS

Weihang Ren*, University of Kentucky
Xiangrong Yin, University of Kentucky

The T-central subspace, introduced by Luo, Li and Yin (2014), allows one to perform sufficient dimension reduction for any statistical functional of interest. We proposed a general estimator using third moment to estimate the T-central subspace and to capture its higher-order trend. We particularly studied mean, density and expectiles with their associated central subspaces; we proposed a new way to estimate the respective central subspace exhaustively. Theoretical results were established and simulation studies showed the advantages of our proposed methods.

email: weihang.ren@uky.edu

122. LONGITUDINAL DATA AND JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

BAYESIAN JOINT MODELING OF NESTED REPEATED MEASURE WITH THE PRESENCE OF INFORMATIVE DROPOUT

Enas Mustfa Ghulam*, University of Cincinnati and Cincinnati Children's Hospital Medical Center
Rhonda D. Szczesniak, Cincinnati Children's Hospital Medical Center

Longitudinal studies play a major role in epidemiology, clinical research, and medical monitoring. Such studies can involve intensive longitudinal collection over different occasions, known as nested repeated measures (NRM). Dropout is a common challenge in longitudinal studies. If the missingness induced by dropout depends on the unobserved response at or after the time of dropout, then this is called informative dropout. Joint Models are widely used for analyzing longitudinal data with informative dropout. This is the first study to our knowledge that propose a Bayesian parametric joint model for NRM data with informative dropout. We will illustrate the proposed methodology with real data collected from a prospective ambulatory blood pressure monitoring study. The original aim of the study was to characterize rate of change in diastolic blood pressure from children with obstructive sleep apnea after surgical intervention, and compare it to levels arising from healthy controls. However, over 60% of children were lost to follow up during the study before completing the measurement schedule.

email: ghulamem@mail.uc.edu

THE JOINT MODELLING OF LONGITUDINAL PROCESS AND CENSORED QUANTILE REGRESSION

Bo Hu*, Columbia University
Ying Wei, Columbia University
Mary Beth Terry, Columbia University

Censored quantile regression is a flexible survival model that complements traditional likelihood based approaches, such as the Cox model. Conditional quantiles can model a survival outcome without pre-specifying a parametric likelihood function. Existing censored quantile methods are mostly limited to fixed cross-sectional covariates, while in many longitudinal studies, researchers wish to investigate the associations between longitudinal covariates and a survival outcome. We propose a framework that jointly model a longitudinal covariate process, conditional quantiles of a survival outcome and their associations. This framework is an extension of a censored quantile based data augmentation algorithm (He et al., 2018), to allow for a longitudinal covariate process in presence of a mixture of censoring schemes. We apply the proposed method to the LEGACY Girls Study to understand the influence of individual genetic profiles on the pubertal development while adjusting for BMI growth trajectories. The results are compared favorably to likelihood based joint modeling approaches and yield new insight on the puberty growth of girls with breast cancer family history.

email: bh2567@columbia.edu

JOINT LATENT CLASS TREES: A TREE-BASED APPROACH TO JOINT MODELING OF TIME-TO-EVENT AND LONGITUDINAL DATA

Jeffrey S. Simonoff*, New York University
Ningshan Zhang, New York University

Joint modeling of longitudinal and time-to-event data provides insights into the association between the two quantities. The joint latent class modeling approach assumes that conditioning on latent class membership, the trajectories of longitudinal data are independent of survival risks. The resulting latent classes provide a data-dependent clustering of the population, which is also of interest. The most common parametric approach, the joint latent class model (JLCM), further restricts analysis to using time-invariant covariates in modeling survival risks and latent class memberships. We propose a nonparametric joint latent class modeling approach based on trees (JLCT). JLCT is fast to fit, and can use time-varying covariates in all of its modeling components. We compare JLCT with JLCM on simulated data, and demonstrate the prognostic value of using time-varying covariates in each of the modeling components. We further apply JLCT to a real application, and demonstrate again that JLCT admits improved prediction performance, while being orders of magnitudes faster than the parametric approach.

email: jsimonof@stern.nyu.edu

FUSION LEARNING IN STRATIFIED MODELS BY PENALIZED GENERALIZE ESTIMATING EQUATIONS

Lu Tang*, University of Pittsburgh
Peter X.K. Song, University of Michigan

Stratification is a commonly used technique in medical studies to handle heterogeneity. Two issues of stratification-based strategies are (i) whether individual strata are sufficiently distinctive to warrant stratification, and (ii) sample size attrition resulted from the stratification leads to loss of statistical power. To overcome these issues, we propose a penalized generalized estimating equations approach to identifying homogeneous subgroups of parameters in correlated data analysis where we allow a data-driven flexible stratification. Specifically, we develop a fusion approach that identifies homogeneous parameters from strata-specific models according to the similarity of their estimates, so that we achieve both adequate stratification and larger sample sizes in individual strata. The proposed approach also allows us to assess whether or not, if so how many, strata are needed to build the stratified model. The method is evaluated by numerical studies and applied to a psychiatric study with nonignorable missing data.

email: lutang@pitt.edu

MIXTURE OF LINEAR MIXED EFFECTS MODELS WITH REAL DATA APPLICATION

Yian Zhang*, New York University
Lei Yang, New York University
Zhaoyin Zhu, New York University
Yongzhao Shao, New York University

Linear mixed effects (LME) models are commonly used in medical studies and other applications for longitudinal data. In reality, these data can be from a mixture of different groups of subjects. Then, it is preferred to fit a group-specific model if the group indicator is available. In this study, we consider a case that longitudinal data come from two groups of subjects but the group indicator is only available for a part of subjects. We construct a finite mixture of LME models via joint modeling of missing group indicator using a logit model and longitudinal data using LME models. We propose a penalized likelihood method with adaptive LASSO penalty using EM algorithm for variable selection in this mixture model. Simulation studies are used to demonstrate the performance of the proposed method. The ADNI (https://ida.loni.usc.edu/) dataset is used to illustrate utility of the introduced model in selecting factors that relate to the latent groups with differential conversion from mild cognitive impairment (MCI) to Alzheimer's disease (AD) and in identifying factors that associate to longitudinal cognitive scores for subjects in each of the subgroups simultaneously.

email: yz2777@nyu.edu

AN APPROXIMATE APPROACH FOR FITTING TWO-PART MIXED EFFECTS MODELS TO LONGITUDINAL SEMI-CONTINUOUS DATA

Hyoyoung Choo-Wosoba*, National Cancer Institute, National Institutes of Health
Paul S. Albert, National Cancer Institute, National Institutes of Health

Two-part mixed effects models have been widely used in which the binary and continuous components are modeled separately with random effects that are correlated between the two components. One of the main challenges for likelihood-based approaches for these models is that it requires numerical integration over a potentially large number of correlated random effects. For this reason, no existing software is available for fitting these models for random effect structures other than for two correlated random intercepts. We propose an imputation approach that will allow practitioners to separately use standard linear and generalized linear mixed models to estimate the fixed effects for two-part mixed effects models with complex random effects structures. We apply a conditional approximation approach to impute positive continuous values corresponding to zero measurements. Specifically, we show that for a wide range of parameter values, our proposed approach results in nearly unbiased estimates and is simple to implement with standard software. We illustrate the proposed imputation approach for the analysis of longitudinal clinical trial data with many zeros.

email: hyoyoung.choo-wosoba@nih.gov

MONITORING PROGRESSION TOWARDS RENAL FAILURE: LESSONS FROM A LARGE VA COHORT

Fridtjof Thomas*, University of Tennessee Health Science Center
Oguz Akbilgic, University of Tennessee Health Science Center
Praveen K. Potukuchi, University of Tennessee Health Science Center
Keiichi Sumida, University of Tennessee Health Science Center
Csaba P. Kovesdy, Memphis VA Medical Center

End stage renal disease (ESRD) is the last stage of chronic kidney disease (CKD) and progression towards renal failure is gauged by the estimated glomerular filtration rate (eGFR), especially whether any eGFR decline accelerates beyond what is expected by healthy aging. We use a cohort of over 3 million US veterans of which 363,021 developed incident CKD and 10,215 ESRD and fit a longitudinal mixed-effects model with an added continuous-time, non-stationary stochastic process representing the change of an individual patient's eGFR as suggested by Diggle et al. (2015). From this model, we derive the individual predictive probabilities that a patient's respective true underlying rate of decline in GFR exceeds 5% (as used by Diggle et al.) at any point in time and develop evidence based expectations for the sensitivity and specificity of decision rules possibly based on this approach of eGFR monitoring with respect to predicting incidence CKD and/or ESRD. Reference: Diggle, P.J., I. Sousa, and Ö. Asar, Real-time monitoring of progression towards renal failure in primary care patients. Biostatistics, 2015. 16(3): p. 522-536.

email: fthomas4@uthsc.edu
