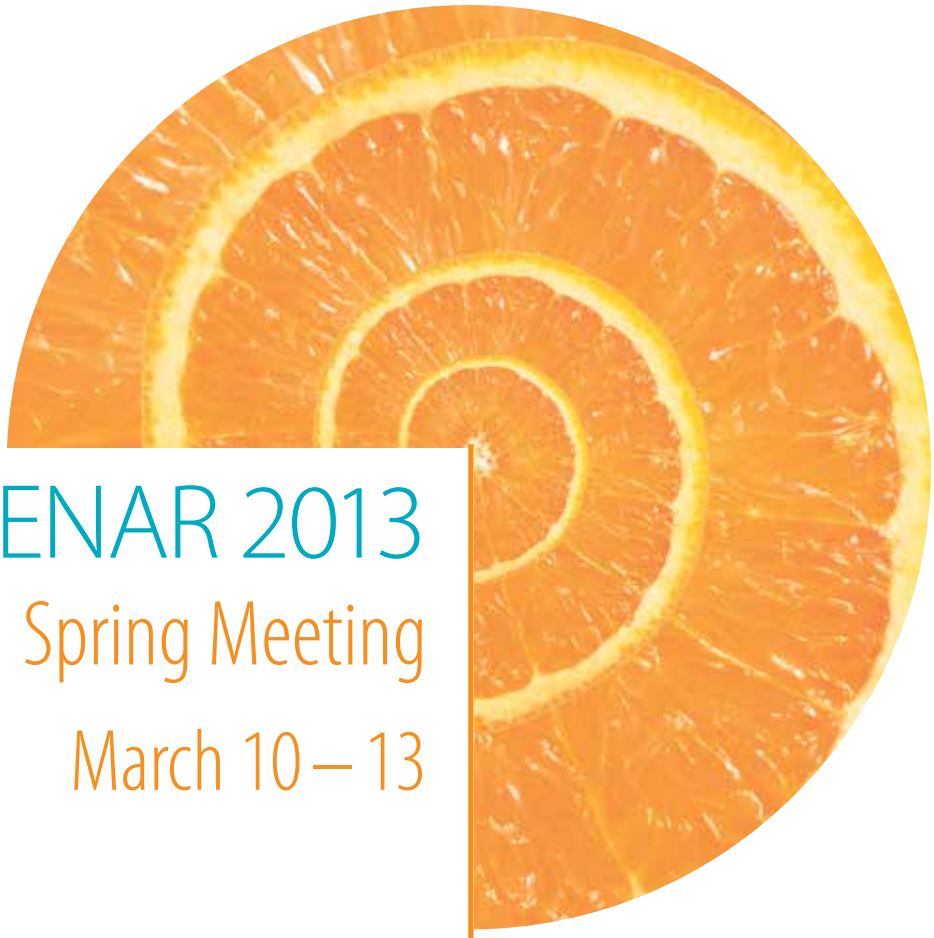
Abstracts



ENAR 2013

Spring Meeting

March 10 – 13

With IMS and
Sections of ASA

ENAR

Orlando World Center Marriott Resort | Orlando, Florida

STATISTICS2013.ORG
STATISTICS

1. POSTERS: CLINICAL TRIALS AND STUDY DESIGN

1a. OPTIMAL BAYESIAN ADAPTIVE TRIAL OF PERSONALIZED MEDICINE IN CANCER

Yifan Zhang*, Harvard University
Lorenzo Trippa, Harvard University, Dana-Farber Cancer Institute
Giovanni Parmigiani, Harvard University, Dana-Farber Cancer Institute

Clinical biomarkers play an important role in personalized medicine in cancer clinical trials. An adaptive trial design enables researchers to use treatment results observed from early patients to aid in treatment decisions of later patients. We describe a biomarker-incorporated Bayesian adaptive trial design. This trial design is the optimal strategy that maximizes the total patient responses. We study the effects of the biomarker and marker group proportions on the total utility and present comparisons between the optimal trial design with other adaptive trial designs.

email: yifanzhangyifan@gmail.com

1b. INTERACTIVE Q-LEARNING FOR DYNAMIC TREATMENT REGIMES

Kristin A. Linn*, North Carolina State University
Eric B. Laber, North Carolina State University
Leonard A. Stefanski, North Carolina State University

Forming evidence-based rules for optimal treatment allocation over time is a priority in personalized medicine research. Such rules must be estimated from data collected in observational or randomized studies. Popular methods for estimating optimal sequential decision rules from data, such as Q-learning, are approximate dynamic programming algorithms that require modeling non-smooth transformations of the data. Postulating a simple, well-fitting model for the transformed data can be difficult, and under many simple generative models the most commonly employed working models—namely linear models—are known to be misspecified. We propose an alternative strategy for estimating optimal sequential decision rules wherein all modeling takes place before applying non-smooth transformations of the data. This simple change of ordering between modeling and transforming the data leads to high quality estimated sequential decision rules. Additionally, the proposed estimators involve only conditional mean and variance modeling of smooth functionals of the data. Consequently, standard statistical procedures for exploratory analysis, model building, and validation can be used. Furthermore, under minimal assumptions, the proposed estimators enjoy simple normal limit theory.

email: kalinn@ncsu.edu

1c. SEMIPARAMETRIC PROPORTIONAL RATE REGRESSION FOR THE COMPOSITE ENDPOINT OF RECURRENT AND TERMINAL EVENTS

Lu Mao*, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill

Analysis of recurrent event data has received tremendous attention in recent years. A major complication arises when recurrent events are terminated by death. To assess the overall covariate effects, we consider the composite endpoint of recurrent and terminal events and propose a proportional rate model which specifies that (possibly time-dependent) covariates have multiplicative affects on the marginal rate function of the composite event process. We derive appropriate estimators for the regression parameters and the baseline mean function by modifying the familiar inverse probability weighting technique. We show that the estimators are consistent and asymptotically normal with variances that can be consistently estimated. Simulation studies demonstrate that the proposed methods perform well in realistic situations. An application to the Community Programs for Clinical Research on AIDS (CPCRA) study is provided.

email: lmao@unc.edu

1d. DETECTION OF OUTLIERS AND INFLUENTIAL POINTS IN MULTIVARIATE LONGITUDINAL MODELS

Yun Ling*, University of Pittsburgh School of Medicine
Stewart J. Anderson, University of Pittsburgh School of Medicine
Richard A. Bilonick, University of Pittsburgh School of Medicine
Gadi Wollstein, University of Pittsburgh School of Medicine

In clinical trials, multiple characteristics of individuals are repeatedly measured. Multivariate longitudinal data allow one to analyze the joint evolution of multiple characteristics over time. Detection of outlier and influential points for multivariate longitudinal data is important to understand potentially critical multivariate observations which can unduly influence the results of analyses. In this presentation, we propose a new approach that extends Cook's distance to multivariate mixed effect models, conditional on different characteristics and subjects. Our approach allows different types of outliers and influential points: it could be one or more measurements on an individual at a single time point, or all measurements on that individual over time. Our approach also takes

into account (1) different residual variances for different characteristics; (2) the correlation among residuals for the same characteristic measured at different time points (within-characteristic correlation); (3) the correlation among residuals for different characteristics measured at one time point (inter-characteristic correlation); and (4) unequally spaced assessments where not all responses of different individuals are measured at the same time points. We apply the approach to the analysis of retina data for glaucoma eyes from UPMC.

email: yul27@pitt.edu

1e. TESTS FOR EQUIVALENCE OF TWO SURVIVAL FUNCTIONS IN PROPORTIONAL ODDS MODEL

Elvis Martinez*, Florida State University
Debajyoti Sinha, Florida State University
Wenting Wang, Florida State University
Stuart R Lipsitz, Harvard University

When survival responses from two treatment arms satisfy the proportional odds survival models (POSM), we present a proper statistical formulation of the clinical hypothesis of therapeutic equivalence. We show that difference between two survival functions being within maximum allowable difference, implies the survival odds parameter to be within a specific interval and vice versa. Our equivalence test, formulation, and related procedure are applicable even in presence of additional covariates beyond treatment arms. Our theoretical and simulation studies show that actual type I error rate for popular equivalence testing procedure (Wellek, 1993) under proportional hazards model (PHM) is higher than the intended nominal rate when the true model is POSM. Whereas, our POSM based procedures have correct type I error rates under the POSM as well as PHM. These investigations show that instead of using log-rank based tests, repeated use of our test will be a safer statistical practice for equivalence trials of survival responses.

email: elvism@stat.fsu.edu

1f. A CLASS OF IMPROVED HYBRID HOCHBERG-HOMMEL TYPE STEP-UP MULTIPLE TEST PROCEDURES

Jiangtao Gou*, Northwestern University
Dror Rom, Prosoft, Inc.
Ajit C. Tamhane, Northwestern University
Dong Xi, Northwestern University

The p-value based step-up multiple test procedure of Hochberg (1988) is very popular in practice because of its simplicity and because it is more powerful than the equally simple to use the step-down procedure of Holm (1979). The Hommel (1988) procedure is even more powerful than the Hochberg procedure but it is less widely used because it is not as simple and is less intuitive. In this paper we derive a new procedure which improves upon the Hommel procedure by gaining power as well as having a simple step-up structure similar to the Hochberg procedure. Thus it offers the best choice among all p-value based stepwise multiple test procedures. The key to this improvement is employing a consonant procedure whereas the Hommel procedure is not consonant and can be improved (Romano, Shaikh, and Wolf, 2011). Exact critical constants of this new procedure can be numerically calculated and tabled. But the 0th order approximations to the exact critical constants, albeit slightly conservative, are simple to use and need no tabling, and hence are recommended in practice. Adjusted p-values of this proposed procedure are derived. The proposed procedure is shown to control the familywise error rate (FWER) both under independence (analytically) and dependence (via simulation) among the p-values, and also shown to be more powerful (via simulation) than competing procedures. Illustrative examples are given.

email: jgou@u.northwestern.edu

1g. CHARACTERIZATION OF TWO-STAGE CONTINUAL REASSESSMENT METHOD FOR DOSE FINDING CLINICAL TRIALS

Xiaoyu Jia, Columbia University

The continual reassessment method (CRM) is an increasingly popular model-based method for dose finding clinical trials among clinicians. A common practice is to use the CRM in a two-stage design, whereby the model-based CRM is activated only after an initial sequence of patients are tested. While there are practical appeals of the two-stage CRM approach, the theoretical framework is lacking in the literature. As a result, it is often unclear how the CRM design components (such as the initial dose sequence and the dose toxicity model) can be properly chosen in practice. This paper studies a theoretical framework that characterizes the design components of a two-stage CRM, and proposes a calibration process. A real trial example is used to demonstrate that the proposed process can be implemented in a timely and reproducible manner, and yet offers competitive operating characteristics when compared to a labor-intensive ad hoc calibration process. We also illustrate using the proposed framework that the performance of the CRM is insensitive to the choice of the dose-toxicity model.

xj2119@columbia.edu

2. POSTERS: BAYESIAN METHODS / CAUSAL INFERENCE

2a. BAYESIAN MODELS FOR CENSORED BINOMIAL DATA: RESULTS FROM AN MCMC SAMPLER

Jessica Pruszynski*, Medical College of Wisconsin
John W. Seaman, Jr., Baylor University

Censored binomial data may lead to irregular likelihood functions and problems with statistical inference. In previous studies, we have compared Bayesian and frequentist models and shown that Bayesian models outperform their frequentist counterparts. In this study, we compare the performance of a Bayesian model under varying sample sizes and prior distributions. We include results from a simulation study in which we compare properties such as point estimation, interval coverage, and interval width.

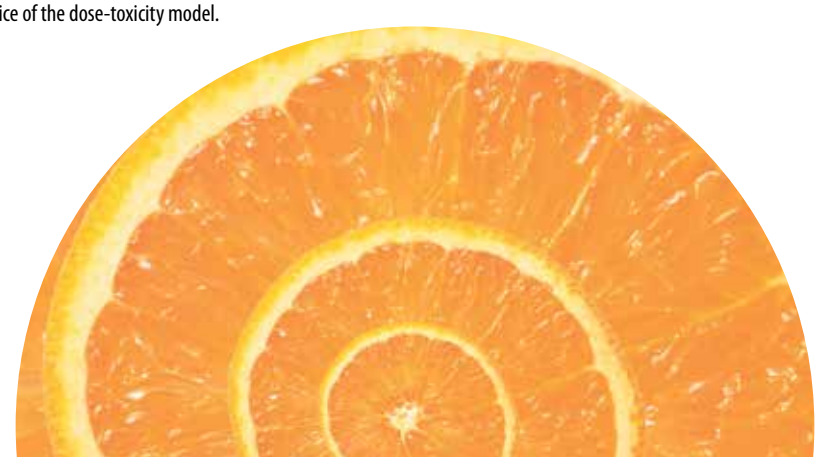
email: jpruszynski@mcw.edu

2b. AN APPROXIMATE UNIFORM SHRINKAGE PRIOR FOR A MULTIVARIATE GENERALIZED LINEAR MIXED MODEL

Hsiang-chun Chen*, Texas A&M University
Thomas E. Wehrly, Texas A&M University

The multivariate generalized linear mixed models (MGLMM) are used for jointly modeling the clustered mixed outcomes obtained when there is more than one response repeatedly measured on each individual in scientific studies. Bayesian methods are one of the most widely used techniques for analyzing MGLMM. The need of noninformative priors arises when there is insufficient information on the model parameters. The main purpose of this study is to propose an approximate uniform shrinkage prior for the random effect variance components in the Bayesian analysis for the MGLMM. This approximate prior is an extension of the approximate uniform shrinkage prior proposed by Natarajan and Kass (2000). This approximate prior is easy to apply and is shown to possess several nice properties. The use of the approximate uniform shrinkage prior is illustrated in terms of both a simulation study and also real world data.

email: ahcchen@stat.tamu.edu



2c. THE BAYESIAN ADAPTIVE BRIDGE

Himel Mallick*, University of Alabama, Birmingham
Nengjun Yi, University of Alabama, Birmingham

We propose the Bayesian adaptive bridge estimator for both general and generalized linear models. The proposed estimator solves the bridge regression model by using a Gibbs sampler. Scale mixture of uniform distribution is considered to yield the MCMC scheme. The proposed method adaptively selects the tuning parameter and selects the concavity parameter from the data. In addition, an ECM algorithm is proposed to estimate the posterior modes of the parameters. Numerical studies and real data analyses are carried forward to compare the performance of the Bayesian adaptive bridge estimator to its frequentist counterpart.

email: himel.stat.iitk@gmail.com

2d. DETECTING LOCAL TWO SAMPLE DIFFERENCES USING DIVIDE-MERGE OPTIONAL POLYA TREES WITH AN APPLICATION IN GENETIC ASSOCIATION STUDIES

Jacopo Soriano*, Duke University
Li Ma, Duke University

Testing if two samples come from the same distribution and identifying local differences is challenging in high dimensional settings. We propose a new Bayesian nonparametric approach to address these problems. We introduce a prior called the DIvide-MErge Optional Polya Tree (dime-OPT), which is constructed based on an optional Poya tree (OPT) process that can split into multiple OPTs on subsets of the sample space where local difference exists between the sample distributions, and can permanently merge into a single process on parts of the space where the sample distributions are conditionally identical. We show that two-sample tests based on the posterior of this flexible prior achieve higher power than existing methods for detecting differences lying in small regions of the space. Moreover, we suggest methods to identify and visualize where the difference is located. For illustrative purposes this method is applied to a genetic association study.

email: jacopo.soriano@duke.edu

2e. EFFICIENT BAYESIAN QUANTITATIVE TRAIT LOCI (QTL) MAPPING FOR LONGITUDINAL TRAITS

Wonil Chung*, University of North Carolina, Chapel Hill
Fei Zou, University of North Carolina, Chapel Hill

In this paper, we extend the Bayesian QTL mapping model with a composite model space framework to handle longitudinal traits. To further improve the flexibility of the Bayesian model, we propose a grid-based method to model the covariance structure of the data, which can accurately approximate any complex covariate structure. The dimension of the working covariance matrix depends on the number of fixed grid points even though each subject may have different number of measurements at different time points. We apply Chen and Dunson's method with a modified Cholesky decomposition to estimate the covariance of random effect. In addition, the proposed method jointly models the main and interactions of all candidate SNPs jointly. It can improve power for mapping genes interacting with other genes or non-genetic factors. The simulation study shows that proposed Bayesian method with all time points outperformed ordinary Bayesian method with one time point. We analyzed GAW18 blood pressure data and identified SNPs with suggestive evidence on Chromosome 1, 3, 7 and 15 for systolic blood pressure (SBP), and SNPs on chromosome 19 for diastolic blood pressure (DBP).

email: wchung11@live.unc.edu

2f. HIERARCHICAL BAYESIAN MODEL FOR COMBINING INFORMATION FROM MULTIPLE BIOLOGICAL SAMPLES WITH MEASUREMENT ERRORS: AN APPLICATION TO CHILDREN PNEUMONIA ETIOLOGY STUDY

Zhenke Wu*, Johns Hopkins Bloomberg School of Public Health
Scott L. Zeger, Johns Hopkins Bloomberg School of Public Health

Determining the etiology of hospitalized severe and very severe pneumonia is difficult because of the absence of a single test that is both highly sensitive and specific. As a result, The Pneumonia Etiology Research for Childhood Pneumonia (PERCH) study collects multiple specimens and performs multiple tests on those specimens, each with varying sensitivity and specificity. The volume and complexity of data present an analytic challenge to identifying the pathogen causing infection. Current methods are mainly model-free and rule-based; they do not systematically incorporate the major sources of information and statistical uncertainties. We propose a statistical model to estimate the prevalence of infection for each pathogen among cases, and the attributable risk of each of these pathogens as a measure of the putative cause of pneumonia. The advantage of this approach is that it allows for multiple imprecise measures of each pathogen from multiple biological samples. The method naturally combines the multiple measures into an optimal composite measure with reduced measurement error. We develop and apply a nested multinomial model that incorporates the natural biological hierarchy of pathogens for efficient computation. Extensions of our current model to include misclassification of pneumonia case status and quantitative pathogen level measurements are also discussed.

email: zhwu@jhsph.edu

2g. A BAYESIAN MISSING DATA FRAMEWORK FOR GENERALIZED MULTIPLE OUTCOME MIXED TREATMENT COMPARISONS

Hwanhee Hong*, University of Minnesota
Haitao Chu, University of Minnesota
Jing Zhang, University of Minnesota
Robert L. Kane, University of Minnesota
Bradley P. Carlin, University of Minnesota

Bayesian statistical approaches to mixed treatment comparisons (MTCs) are becoming more popular due to their flexibility and interpretability. Many randomized clinical trials report multiple outcomes with possible inherent correlations. Moreover, MTC data are typically sparse and researchers often choose study arms based on previous trials. In this paper, we summarize existing hierarchical Bayesian methods for MTCs with a single outcome, and we introduce novel Bayesian approaches for multiple outcomes simultaneously, rather than in separate MTC analyses. We do this by incorporating missing data and correlation structure between outcomes through contrast- and arm-based parameterizations that consider any unobserved treatment arms as missing data to be imputed. We also extend the model to apply to all types of generalized linear model outcomes, such as count or continuous responses. We develop a new measure of inconsistency under our missing data framework, having more straightforward interpretation and implementation than standard methods. We offer a simulation study under various missingness mechanisms (e.g., MCAR, MAR, and MNAR) providing evidence that our models outperform existing models in terms of bias and MSE, then illustrate our methods with two real MTC datasets. We close with a discussion of our results and a few avenues for future methodological development.

email: hong0362@umn.edu

2h. LARGE SAMPLE RANDOMIZATION INFERENCE OF CAUSAL EFFECTS IN THE PRESENCE OF INTERFERENCE

Lan Liu*, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill

Recently, an increasing amount of attention has focused on making causal inference when interference is possible, i.e., when the potential outcomes of one individual may be affected by the treatment (or exposure) of other individuals. For example, in infectious diseases, whether one individual becomes infected may depend on whether another individual is vaccinated. In the presence of interference, treatment may have several types of effects. In this paper, we consider inference about such effects when the population consists of groups of individuals where interference is possible within groups but not between groups. The asymptotic distributions of estimators of

the causal effects are derived when either the number of individuals per group or the number of groups grows large. A simulation study is presented showing that in various settings the corresponding asymptotic confidence intervals have good coverage in finite samples and are substantially narrower than exact confidence intervals.

email: lanl@email.unc.edu

2i. EFFICIENT SAMPLING METHODS FOR MULTIVARIATE NORMAL AND STUDENT-t DISTRIBUTIONS SUBJECT TO LINEAR CONSTRAINTS

Yifang Li*, North Carolina State University
Sujit K. Ghosh, North Carolina State University

Sampling from a truncated multivariate normal distribution subject to multiple linear inequality constraints is a recurring problem in many areas in statistics and econometrics, such as the order restricted regressions, censored models, and shape-restricted nonparametric regressions. However, this is not an easy problem due to the existence of the normalizing constant involving the probability of the multivariate normal distribution. In this paper, we establish an efficient mixed rejection sampling method for the truncated univariate normal distribution. Our method has uniformly larger acceptance rates than the popular existing methods. Since the full conditional distribution of a truncated multivariate normal distribution is also truncated normal, we employ our univariate sampling method and implement the Gibbs sampler for sampling from the truncated multivariate normal distribution with convex polytope restriction regions. Experiments show that our proposed Gibbs sampler is accurate and has good mixing property with fast convergence. Since a Student-t distribution can be obtained by taking the ratio of a multivariate normal distribution and an independent chi-squared distribution, we can also easily generate this sampling method to the truncated multivariate Student-t distribution, which is also a common encountered problem.

email: yli40@ncsu.edu

3. POSTERS: MICROARRAY ANALYSIS / NEXT GENERATION SEQUENCING

3a. PROFILING CANCER GENOMES FROM MIXTURES OF TUMOR AND NORMAL TISSUE VIA AN INTEGRATED STATISTICAL FRAMEWORK WITH SNP MICROARRAY DATA

Rui Xia*, University of Texas MD Anderson Cancer Center
Selina Vattathil, University of Texas MD Anderson Cancer Center
Paul Scheet, University of Texas MD Anderson Cancer Center

Current methods for inference of somatic DNA aberrations in tumor genomes using SNP DNA microarrays require sufficiently high tumor purities (10-15%) to detect an increase in the variation of particular features of the microarray, such as the B allele frequency or total allelic intensity (logR ratio). By incorporating information from the germline genome, we are able to detect aberrations at vastly lower tumor proportions (e.g. 3-4%). Our likelihood-based approach integrates a hidden Markov model (HMM) for population haplotype variation for the germline genome (Scheet & Stephens, AJHG 78:629, 2006) with an HMM for DNA aberrations in the tumor. Thus, our approach directly accounts for the perturbations in the data that would be expected from actual chromosomal-level aberrations. Our method reports mean allele specific copy number, as well as marginal probabilities of aberration types in tumor DNA. We test our method on real and simulated data based on breast cancer cell line and Illumina 370K array, and identify aberration regions of 11Mb length in 3% tumor purities. We expect our method to provide more accurate inference of copy number changes in a variety of settings (tumor profiles, somatic variation). And more generally, our approach establishes an integrated statistical framework for studying inherited and tumor genomes.

email: rxia@mdanderson.org

3b. A PROFILE-TEST FOR MICRORNA MICROARRAY DATA ANALYSIS

Bin Wang*, University of South Alabama

MicroRNA is a set of small RNA molecules mediating gene expression at post-transcriptional/translational levels. Most of well-established high throughput discovery platforms, such as microarray, real time quantitative PCR, and sequencing, have been adapted to study microRNA in various human diseases. Analyzing microRNA data is challenging due to the fact that the total number of microRNAs in humans small and the majority of microRNA maintains relatively low abundance in the cells. The signals of these low-expressed microRNAs are greatly affected by non-specific signals including the background noise. We studied the microRNA microarrays obtained with multiple platforms, and developed a measurement error model-based test for differentially-expressed microRNA detection, at both probe-level and profile-level.

email: bwang@southalabama.edu

3c. A MULTIPLE TESTING METHOD FOR DETECTING DIFFERENTIALLY EXPRESSED GENES

Linlin Chen*, Rochester Institute of Technology
Alexander Gordon, University of North Carolina, Charlotte

Due to the existence of strong correlations between expression levels of different genes, the procedures which are commonly used to detect the genes differentially expressed between two or more phenotypes are unable to overcome the two main problems: high instability of the number of false discoveries and low power. It may be impossible to completely understand these correlations due to the complexity of their biological nature. We have proposed a new multiple testing method to balance type I and type II errors in an optimal, in a sense, way. However, the correlation structure of microarray data is still the main obstacle standing in the way of this and other gene selection procedures. To remove this obstacle, we further improve the statistical methodology by exploiting the property of low dependency between the terms of the so-called δ-sequence proposed by Klebanov and Yakovlev ($2007b$). In this paper, we will review and further study the application of the δ-sequence in the selection of the significantly changed genes. We will examine the use of the δ-sequence in conjunction with the Bonferroni adjustment and balancing type I and type II errors, and will discuss the results of analysis of some real microarray data. The comparison with the univariate gene selection method will also be discussed.

email: linlin.chen@gmail.com

3d. APPLICATION OF BILINEAR MODELS TO THREE GENOME-WIDE EXPRESSION ANALYSIS PROBLEMS

Pamela J. Lescault, University of Vermont
Julie A. Dragon, University of Vermont
Jeffrey P. Bond*, University of Vermont

The development of biomedical processes that include the production and interpretation of genome-wide expression profiles often involves comparison of alternative technologies. We study two examples in which expression profiles are obtained from each member of a set of cellular samples using two different technologies. In the first case the two technologies are alternatives for obtaining purified cell populations from cell mixtures. In the second case the two technologies are alternatives for obtaining gene expression profiles from RNA samples.

This class of problems is distinguished from many two-way genome-wide expression experiment designs by the need to capture separately the latent variation associated with each technology. Bilinear models, including the Surrogate Variable Analysis of Leek and Storey, allowed us to capture and compare the latent variation associated with each technology.

email: Jeffrey.Bond@uvm.edu

3e. REPRODUCIBILITY OF THE NEUROBLASTOMA GENE TARGET ANALYSIS PLATFORM

Pamela J. Lescault, University of Vermont
Julie A. Dragon*, University of Vermont
Jeffrey P. Bond, University of Vermont
Russ Ingersoll, Intervention Insights
Giselle Sholler, Van Andel Institute

The goal of the Neuroblastoma Gene Target Analysis Platform (NGTAP) is to use RNA expression profiling of neuroblastoma tumors to select drug therapies for patients with recurrent or refractory neuroblastoma. The NGTAP includes a process that, based on fine needle aspiration of a neuroblastoma solid tumor, produces three multivariate observations: 1) an expression profile based on Affymetrix GeneChip technology, 2) a comparative expression profile obtained by centering and scaling the expression profile using the location and standard deviation of a normal whole body tissue set, 3) a set of drug therapies obtained from comparative expression profiles using the Intervention Insights OncInsights service (based on work by Craig Webb at the Van Andel Research Institute). Each of six fine needle aspirates was dissected into three portions. Distance-based nonparametric multivariate analysis of variance allowed us to evaluate, for each of the three types of observations, the variation between patients in the context of the variation within biopsies. The reproducibility averaged over patients, replicates, and drugs is high. Reproducibility increases as the threshold drug score increases.

email: Julie.Dragon@uvm.edu

3f. HIGH DIMENSIONAL EQUIVALENCE TESTING USING SHRINKAGE VARIANCE ESTIMATORS

Jing Qiu, University of Missouri, Columbia
Yue Qi*, University of Missouri, Columbia
Xiangqin Cui, University of Alabama

Identifying differentially expressed genes has been an important and widely used approach to investigate gene functions and molecular mechanisms. A related issue that has drawn much less attention but is equally important is the identification of constantly expressed genes across different conditions. The common practice is to treat genes that are not significantly differentially expressed as significantly equivalently expressed. Such naive practice often leads to large false discovery rate and low power. The more appropriate way for identifying constantly expressed genes should be conducting high dimensional statistical equivalence tests. A well-known equivalence test, the two one-sided tests (TOST), can be used for this purpose. However, due to the "large p and small n" feature of the genomics data, the variance estimator in the TOST test could be unstable. Hence it would be fitting to examine the application of shrinkage variance estimators to the high dimensional equivalence test. In this paper, we aim to apply a shrinkage variance estimator to the TOST test and derive analytic formulas for the p-values of the resultant shrinkage variance equivalence tests. We study the effect of the shrinkage variance estimators on the power of the high dimensional equivalence test through simulation studies and data analysis.

email: yqrx7@mail.missouri.edu

3g. REMOVING BATCH EFFECTS FOR PREDICTION PROBLEMS WITH FROZEN SURROGATE VARIABLE ANALYSIS

Hilary S. Parker*, Johns Hopkins Bloomberg School of Public Health
Hector Corrada Bravo, University of Maryland, College Park
Jeffrey T. Leek, Johns Hopkins Bloomberg School of Public Health

Batch effects are responsible for the failure of promising genomic prognostic signatures, major ambiguities in published genomic results, and retractions of widely-publicized findings. Batch effect corrections have been developed to remove these artifacts, but they are designed to be used in population studies. But genomic technologies are beginning to be used in clinical applications where samples are analyzed one at a time for diagnostic, prognostic, and predictive applications. There are currently no batch correction methods that have been developed specifically for prediction. In this paper, we propose a new method called frozen surrogate variable analysis (fSVA) that borrows strength from a training set for individual sample batch correction. We show that fSVA improves prediction accuracy in simulations and in public genomic studies. fSVA is available as part of the sva Bioconductor package.

email: hiparker@jhsph.edu

3h. FEATURE SELECTION AMONG ORDINAL CLASSES FOR HIGH-THROUGHPUT GENOMIC DATA

Kellie J. Archer*, Virginia Commonwealth University
Andre A.A. Williams, National Jewish Health

For most gene expression microarray datasets where the phenotypic variable of interest is ordinal, the analytical approach taken has been either to perform several dichotomous class comparisons or to collapse the ordinal variable into two categories and perform t-tests for each feature. In this talk we review four different ordinal response methods, namely, the cumulative logit, adjacent category, continuation ratio, and stereotype logit model, as feature selection methods for high-dimensional datasets. We will present the results from our simulation study conducted to compare these four methods to the two dichotomous response approaches. As expected, the Type I errors for the ordinal methods are close to the nominal level, while performing several dichotomous class comparisons yields an inflated Type I error. Additionally, the ordinal methods have higher power when compared to the collapsed category t-test method, regardless of whether or not the proportional odds assumption was met. We further illustrate the application of the aforementioned methods for modeling an ordinal phenotypic variable using a publicly available gene expression dataset from Gene Expression Omnibus. We conclude by describing extensions for such methods to multivariable modeling and ordinal response machine learning methods.

email: kjarcher@vcu.edu

3i. A RANK-BASED REGRESSION FRAMEWORK TO ASSESS THE COVARIATE EFFECTS ON THE REPRODUCIBILITY OF HIGH-THROUGHPUT EXPERIMENTS

Qunhua Li*, The Pennsylvania State University

The outcome of high-throughput experiments is affected by many operating parameters in experimental protocols and data-analytical procedures. Understanding how these factors affect the experimental outcome is critical for designing protocols that produce replicable discoveries. The correspondence curve (Li et al 2011) is a scale-free graphical tool to illustrate how reproducibly top-ranked signals are reported at different levels of significance. Though this method provides the convenience to compare the reproducibility of outcome produced at different operating parameters, without linking the method to any particular decision criterion a succinct summary often is more convenient for comparing the effects of different operating parameters. In this work, we propose a regression framework to incorporate the covariate effects in correspondence curves. This frame-

work allows one to characterize the simultaneous or independent effects of covariates on reproducibility and to compare reproducibility while controlling for potential confounding variables. We illustrate this method using data from ChIP-seq experiments.

email: qunhua.li@psu.edu

3j. NONPARAMETRIC METHODS FOR IDENTIFYING DIFFERENTIAL BINDING REGIONS WITH ChIP-Seq DATA

Qian Wu*, University of Pennsylvania School of Medicine
Kyoung-Jae Won, University of Pennsylvania School of Medicine
Hongzhe Li, University of Pennsylvania School of Medicine

ChIP-Seq provides a powerful method for detecting binding sites of DNA-associated proteins, e.g. transcription factors (TFs) and histone modification marks (HMs). Previous research has focused on developing peak-calling procedures to detect the binding sites for TFs. However, these procedures have difficulty when applied to ChIP data of HMs. In addition, it is also important to identify genes with differential binding regions between two experimental conditions, such as different cellular states or different time points. Parametric methods based on Poisson/Negative Binomial distribution have been proposed to address this problem and most require biological replications. However, many ChIP-Seq data usually have a few or even no replicates. We propose a novel nonparametric method to identify the differential binding regions that can be applied to the ChIP-Seq data of TF or HM, even without replicates. Our method is based on nonparametric hypothesis testing and kernel smoothing. We demonstrate the method using a ChIP-Seq data on comparative epigenomic profiling of adipogenesis of human adipose stromal cells and our method detects nearly 20% of genes with differential binding of HM mark H3K27ac in gene promoter regions. The test statistics also correlate with the gene expression changes well, indicating that the identified differential binding regions are indeed biologically meaningful.

email: wuqian7@gmail.com

3k. IN SILICO POOLING DESIGNS FOR ChIP-Seq CONTROL EXPERIMENTS

Guannan Sun*, University of Wisconsin, Madison
Sunduz Keles, University of Wisconsin, Madison

As next generation sequencing technologies are becoming more economical, large-scale ChIP-seq studies are enabling the investigation of the roles of transcription factor binding and epigenome on phenotypic variation. Studying such variation requires individual level ChIP-seq experiments. Standard designs for ChIP-seq experiments employ a paired control per ChIP-seq sample. Genomic coverage for control experiments is often sacrificed to increase the resources for ChIP samples. However, the quality of ChIP-enriched regions identifiable from a ChIP-seq experiment depends on the quality and the coverage of the control experiments. Insufficient coverage leads to loss of power in detecting enrichment. We investigate the effect of in silico pooling of control samples across biological replicates, treatment conditions, and cell lines across multiple datasets with varying levels of genomic coverage. Our empirical studies that compare in silico pooling designs with the gold standard paired-designs indicate that Pearson correlation of the samples can be used to decide whether or not to perform pooling. Using vast amounts of ENCODE data, we show that pair wise correlations between control samples originating from different biological replicates, treatments, and cell lines can be grouped into two classes representing whether or not in silico pooling leads to power gain in detecting enrichment between the ChIP and the control samples. Our findings have important implications for multiplexing multiple samples.

email: sun@stat.wisc.edu

3l. METHOD FOR CANCELLING NONUNIFORMITY BIAS OF RNA-seq FOR DIFFERENTIAL EXPRESSION ANALYSIS

Guoshuai Cai*, University of Texas MD Anderson Cancer Center
Shoudan Liang, University of Texas MD Anderson Cancer Center

Many biases and effects are inherent in RNA-Seq technology. A number of methods have been proposed to handle these biases and effects in order to accurately analyze differential RNA expression at the gene level. However, to precisely estimate mean and variance by cancelling biases such as those due to random hexamer priming and non-uniformity, modeling at the base pair level is required. We previously showed that the overdispersion rate decreases as sequencing depth increases on the gene level. We tested the hypothesis that the overdispersion rate also decreases as sequencing depth increases on the base pair level. In this study, we found that the overdispersion rate decreased as sequencing depth increased on the base pair level. Also, we found that the influence of local primer sequence on the overdispersion rate was no longer significant after stratification by sequencing depth. In addition, compared with our other proposed models, our beta binomial model with dynamic overdispersion rate was superior. The current study will aid in analysis of RNA-Seq data for detecting and exploring biological problems.

email: GCAI@mdanderson.org

3m. BINARY TRAIT ANALYSIS IN SEQUENCING STUDIES WITH TRAIT-DEPENDENT SAMPLING

Zhengzheng Tang*, University of North Carolina, Chapel Hill
Danyu Lin, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill

In sequencing study, it is a common practice to sequence only the subjects with the extreme values of a quantitative trait. This is a cost-effective strategy to increase power in the association analysis. In the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (ESP), subjects with extremely high or low values of body mass index (BMI), low-density lipoprotein (LDL) or blood pressures (BP) were selected for whole-exome sequencing. For a binary trait of interest, the standard logistic regression-even adjust for the trait of sampling-can give misleading results. We present valid and efficient methods for association analysis under trait-dependent sampling. Our methods properly combine the association results from all studies and more powerful than the standard methods. The validity and efficiency of the proposed methods are demonstrated through extensive simulation studies and ESP real data analysis.

email: ztang@bios.unc.edu

3n. QUANTIFYING COPY NUMBER VARIATIONS USING A HIDDEN MARKOV MODEL WITH INHOMOGENEOUS EMISSION DISTRIBUTIONS

Kenneth McCallum*, Northwestern University
Ji-Ping Wang, Northwestern University

Copy number variations (CNVs) are a significant source of genetic variation and have been found frequently associated with diseases such as cancers and autism. High-throughput sequencing data is increasingly being used to detect and quantify CNVs; however, the distributional properties of the data are not fully understood. A hidden Markov model is proposed using inhomogeneous emission distributions based on negative binomial regression to account for sequencing biases. The model is tested on whole genome sequencing data and simulated data sets. The model based on negative binomial regression is shown to provide a good fit to the data and provides competitive performance compared to methods based on normalization of read counts.

email: kennethmccallum2013@u.northwestern.edu

4. POSTERS: STATISTICAL GENETICS / GENOMICS

4a. DETECTING RARE AND COMMON VARIANT IN NEXT GENERATION SEQUENCING DATA USING A BAYESIAN VARIABLE SELECTION

Cheongeun Oh*, New York University

With the advent of next-generation sequencing, rare variants with a minor allele frequency (MAF) < 1 ~ 5% are getting more attention in GWAS to resolve the precise location of the causal variant(s) in complex trait etiology. Despite their importance, testing for associations between rare variants and traits has proven challenging. Most of current statistical methods for genetic association studies have been developed based on underlying common disease common variant assumption. Although several approaches have been proposed for the analysis of rare variants, evaluating the potential impact of rare variants on disease is complicated by their uncommon nature and underpowered due to the low allele frequencies. In this study, we propose a novel Bayesian variable selection to detect associations with both rare and common genetic variants simultaneously for quantitative traits, in which we utilize the group indicator to collapse rare variants within genomic regions. We evaluate the proposed method and compare its performance to existing methods on extensive simulated data under the high dimensional scenario.

email: cceohh@gmail.com

4b. DOMINANCE MODELING FOR GWAS HIT REGIONS WITH GENERALIZED RESAMPLE MODEL AVERAGING

Jeremy A. Sabourin*, University of North Carolina, Chapel Hill
Andrew Nobel, University of North Carolina, Chapel Hill
William Valdar, University of North Carolina, Chapel Hill

The analysis of complex traits in humans has primarily concentrated on genetic models which most often assume additive only SNP effects. When non-additive effects such as dominance or overdominance are present, additive-only models can be underpowered. We present RMA-dawg, a generalized resample model averaging (RMA) based method using the group LASSO that allows for additive and non-additive SNP effects. RMA-dawg estimates for each SNP, the probability that it would be included in a multiple SNP model in alternative realizations of the data. In modeling dominance, we expose weaknesses of subsampling-based approaches (such as those based on Stability Selection) when considering rare predictors and present a grouping framework that is useful when modeling effects that require multiple predictors per locus, such as dominance. We show that under simulations based on real GWAS data, that RMA-dawg identifies a set of candidates that is enriched for causal loci relative to single locus analysis and standard RMA.

email: jsabouri@unc.edu

4c. EMPIRICAL BAYES ANALYSIS OF RNA-seq WITHOUT REPLICATES FOR MULTIPLE CONDITIONS

Xiaoxing Cheng*, Brown University
Zhijin Wu, Brown University

Recent developments in sequencing technology have led to a rapid increase in RNA sequencing (RNA-seq) data. Identifying differential expression (DE) remains a key task in functional genomics. As RNA-seq application extends to non-model organisms and experiments involving a variety of conditions from environmental samples the need for identifying interesting target RNA-seq without replicates is increasing. We present an empirical Bayes method of estimating differential expression from multiple samples without replicates. There have been a number of statistical methods for the detection of DE in RNA-seq data, which all focus on statistical significance of DE. We argue that, as transcription is an inherently stochastic phenomenon and organisms respond to a lot of environmental cues, it is possible that all expressed genes have DE, only to a different extent. Thus we focus on estimating the magnitude of DE. In our model, each gene is allowed to have DE, but the magnitude of DE in each treatment group is a random variable. So we assume most genes have small differences, the prior for DE is centered at zero. And Maximum a Posteriori (MAP) estimation for the magnitude of differential expression provided is shrunk towards zero.

email: xiaoxing.cheng@gmail.com

4d. ESTIMATING THE NUCLEOTIDE SUBSTITUTION MATRIX USING A FULL FOUR-STATE TRANSITION RATE MATRIX

Ho-Lan Peng*, University of Texas School of Public Health
Andrew R. Aschenbrenner, University of Texas School of Public Health

The nucleotide substitution rate matrix, Q, has been the subject of great importance in molecular evolution because it describes the rates of evolutionary changes across a number of DNA sequences. Historically, the substitution process is assumed to be a continuous time Markov process and has been used to derive probabilities. These probabilities are functions of the parameters and estimation of the parameters is done separately. Our approach is to develop the probabilities in the four state continuous time Markov process with no assumptions on the infinitesimal matrix (giving twelve parameters) and estimate the parameters using maximum likelihood. We will present two methods: a non-homogenous differential equation approach and a spectral decomposition approach. Both methods estimate similar Q and achieve a better model fit than the classical models. We will also compare these methods to the classical models through a simulation.

email: glenn73831@gmail.com

4e. A NETWORK-BASED PENALIZED REGRESSION METHOD WITH APPLICATION TO GENOMIC DATA

Sunkyung Kim*, Centers for Disease Control and Prevention (CDC)
Wei Pan, University of Minnesota
Xiaotong Shen, University of Minnesota

Penalized regression approaches are attractive in dealing with high-dimensional data such as arising in high-throughput genomic studies. New methods have been introduced to utilize the network structure of predictors, e.g. gene networks, to improve parameter estimation and variable selection. All the existing network-based penalized methods are based on an assumption that parameters, e.g. regression coefficients, of neighboring nodes in a network are close in magnitude, which however may not hold. Here we propose a novel penalized regression method based on a weaker prior assumption that the parameters of neighboring nodes in a network are likely to be zero (or non-zero) at the same time, regardless of their specific magnitudes. We propose a novel non-convex penalty function to incorporate this prior, and an algorithm based on difference convex programming. We use simulated data and a gene expression dataset to demonstrate the advantages of the proposed method over some existing methods. Our proposed methods can be applied to more general problems for group variable selection.

email: kimx803@gmail.com

4f. HIERARCHAL MODEL FOR DETECTING DIFFERENTIALLY METHYLATED LOCI WITH NEXT GENERATION SEQUENCING

Hongyan Xu*, Georgia Health Sciences University
Varghese George, Georgia Health Sciences University

Epigenetic changes, especially DNA methylation at CpG loci has important implications in cancer and other complex diseases. With the development of next-generation sequencing (NGS), it is feasible to generate data to interrogate the difference in methylation status for genome-wide loci using case-control design. However, a proper and efficient statistical test is lacking. In this study, we propose a hierarchical model for methylation

data from NGS to detect differentially methylated loci. Simulations under several distributions for the measured methylation levels show that the proposed method is robust and flexible. It has good power and is computationally efficient. Finally, we apply the test to our NGS data on chronic lymphocytic leukemia. The results indicate that it is a promising and practical test.

email: hxu@georgiahealth.edu

4g. MIXED MODELING AND SAMPLE SIZE CALCULATIONS FOR IDENTIFYING HOUSEKEEPING GENES IN RT-PCR DATA

Hongying Dai*, Children's Mercy Hospital
Richard Charnigo, University of Kentucky
Carrie Vyhlidal, Children's Mercy Hospital
Bridgette Jones, Children's Mercy Hospital
Madhusudan Bhandary, Columbus State University

Normalization of gene expression data using internal control genes that have biologically stable expression levels is an important process for analyzing RT-PCR data. We propose a three-way linear mixed-effects model (LMM) to select optimal housekeeping genes. The LMM can accommodate multiple continuous and/or categorical moderator variables with sample random effects, gene fixed effects, systematic effects, and gene by systematic effect interactions. Global hypothesis testing is proposed to ensure that selected housekeeping genes are free of systematic effects or gene by systematic effect interactions. Sample size calculation based on the estimation accuracy of the stability measure is offered to help practitioners design experiments to identify housekeeping genes. We compare our methods with geNorm and NormFinder using case studies.

email: hdai@cmh.edu

4h. LIKELIHOOD BASED INFERENCE ON PHYLOGENETIC TREES WITH APPLICATIONS TO METAGENOMICS

Xiaojuan Hao*, University of Nebraska
Dong Wang, University of Nebraska

Interaction dynamics between the micobiota and the host have taken on ever increasing importance and are now frequently studied. But the statistical methodology for hypothesis testing regarding microbiota structure and composition is still quite limited. Usually, the relative enrichment of one or more taxa is demonstrated with contingency tables formed in an ad hoc manner. To provide a formal statistical framework, we have proposed a likelihood based approach in which the likelihood function is specified in the same manner as for most phylogenetic models using continuous Markov processes with variables representing microbiota structure and composition rather than nucleotide sequences. The computation is performed through a pruning algorithm nested inside a gradient descent based method for parameter optimization. With the availability of a likelihood function, likelihood based inference such as likelihood ratio test can be used to formally test hypothesis of interest. We illustrated the application of this method with a data set pertaining microbial communities in the rat digestive tract under different diet regiments. Different hypotheses regarding microbial community structure were studied with the proposed approach.

email: hxjhelon@gmail.com

4i. A HIGH DIMENSIONAL VARIABLE SELECTION APPROACH USING TREE-BASED MODEL AVERAGING WITH APPLICATION TO SNP DATA

Sharmistha Guha*, University of Minnesota
Saonli Basu, University of Minnesota

As opposed to the existing methodology in high dimensional regression problems, we propose a novel tree-based variable selection approach. Our proposed approach combines different low-rank models, together with model averaging techniques, to yield a model that exhibits far less computational time and greater flexibility in terms of estimation. Simulation examples show high power for the proposed method. We compare our method to some of the current existing methods and show empirically better performance. The proposed approach has been validated using SNP data.

email: sharmistha84@gmail.com

4j. COMPARISON OF STATISTICS IN ASSOCIATION TESTS OF GENETIC MARKERS FOR SURVIVAL OUTCOMES

Franco Mendolia*, Medical College of Wisconsin
John P. Klein, Medical College of Wisconsin
Effie W. Petersdorf, Fred Hutchinson Cancer Research Center
Mari Malkki, Fred Hutchinson Cancer Research Center
Tao Wang, Medical College of Wisconsin

In genetic association studies, there is a need for computationally efficient statistical methods to handle the large number of tests of genetic markers. In this study, we explore several tests based on the Cox proportional hazards models for survival outcomes. We examine the classical partial likelihood-based Wald and score tests and we propose a score statistic which is motivated by Cox-Snell residuals to assess the effects of genetic markers. Computational efficiency and incorporation of these three statistics into a permutation procedure to adjust for multiple testing is addressed. We also consider a simulation-based chi-square test as proposed by Lin (2005) to adjust for multiple testing. Comparison of these four statistics in terms of type I error, power, family-wise error rate and computational efficiency under various scenarios are examined via extensive simulations.

email: fmendolia@mcw.edu

4k. SPARSE MULTIVARIATE FACTOR REGRESSION MODELS AND ITS APPLICATION TO HIGH-THROUGHPUT ARRAY DATA ANALYSIS

Yan Zhou*, University of Michigan
Peter X.K. Song, University of Michigan
Ji Zhu, University of Michigan

The sparse multivariate regression model is a useful tool to explore complex associations between multiple response variables and multiple predictors. When those multiple responses are strongly correlated, ignoring such dependency will impair statistical power and accuracy in the association analysis. In this paper, we propose a new methodology -- sparse multivariate regression factor model (sMFRM), which accounts for correlations of the response variables via lower numbers of unobserved random quantities. This proposed method not only allows us to address the issue that the number of association parameters is much larger than the sample size, but also to account for some unobserved factors that potentially obscure real response-predictor associations. The proposed sMFRM is efficiently implemented by utilizing merits of both the EM algorithm and the group-wise coordinate descend algorithm. The proposed methodology is evaluated through extensive simulation studies. It is shown that our proposed sMFRM outperforms the existing methods in terms of high sensitivity and accuracy in mapping the underlying response-predictor associations. Throughout this paper, we motivate and apply our method with the objective of constructing genetic association networks. Finally, we analyze a breast cancer data adjusting for unobserved non-genetic factors.

email: zhouyan@umich.edu

4l. BAYESIAN GROUP MCMC

Alan B. Lenarcic*, University of North Carolina, Chapel Hill
William Valdar, University of North Carolina, Chapel Hill

In mouse experiments, SNP derived genetics sequences are often encoded and imputed into a sequence of probabilities representing haplotype group membership (i.e. Black6-derived, Mouse-Castenious-derived, etc.) rather than 0-1 SNPs, which are identified through a hidden markov model, such as with the Happy algorithm (Mott et al. 2000). Multi-SNP regression must then follow a mixed-effects framework, with regression coefficients learned at a single SNP representing a set of factors. We implement a sparse Bayesian MCMC framework, built around dynamic memory structures initially designed

for Coordinate Descent Lasso, recording sparse MCMC chains in a compressed sequences, and achieving mode-escape tempering through Equi-Energy sampling (Kou and Wong 2006). Furthermore, we implement a group selection sampler which samples group inclusion based upon a new scheme for sampling between bounding functions. We show that this new sampling mechanism has exponential convergence for all values of inclusion and requires no adaptive sampling. We demonstrate the benefits of group selection in the setting of diallel F1 inheritance experiments for blood pressure and weight gain, as well as 0 (10K) SNP sets from advanced intercross designs.

email: alenarc@med.unc.edu

4m. COMBINING PEPTIDE INTENSITIES TO ESTIMATE PROTEIN ABUNDANCE

Jia Kang*, Merck
Francisco Dieguez, Merck

In proteomics studies, differentially expressed features are often summarized at the peptide level. However, protein level inference is often of great interest to the scientists in order to derive a better understanding of the drug mechanism of action, and to facilitate data integration across multiple platforms (e.g. mRNA profiling). Traditional methods in combining peptide intensities include per-peptide model, per-protein averaging model, and ANOVA model. And it was demonstrated that the ANOVA model outperformed the other two approaches. The ANOVA method, however, is not designed to handle the heteroscedasticity problems encountered in proteomics studies. In this study, we propose a Bayesian approach to aggregate peptide intensities at the protein level, and compare our method to the existing methods through both simulation studies and real data. Our results show that the Bayesian method outperformed the other methods we investigated.

email: jia.kang@merck.com

4n. BOOTSTRAP METHODS FOR GENETIC ASSOCIATION ANALYSIS ON INTERMEDIATE PHENOTYPES IN CASE-CONTROL STUDIES

Naomi Brownstein, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Gary Slade, University of North Carolina, Chapel Hill
Shad Smith, University of North Carolina, Chapel Hill
Luda Diatchenko, University of North Carolina, Chapel Hill
Eric Bair*, University of North Carolina, Chapel Hill

In a case-control study, one may wish to examine associations between putative risk factors and intermediate phenotypes that were ascertained in the study. This is especially common in modern genetic studies, such as genome-wide association studies or next-generation sequencing studies, where the cost of collecting the data is high. Since a case-control study is not a random sample from the population, standard tests for association between intermediate phenotypes and the genetic risk factors will be biased unless appropriate adjustments are performed. There are existing methods for producing unbiased association estimates in this situation, but most existing methods assume that the intermediate phenotype is either binary or continuous and cannot be easily applied to more complicated phenotypes, such as ordinal or survival outcomes. We describe a bootstrap-based method that can produce unbiased estimates of any arbitrary test statistic/regression coefficient in case-control studies. In particular, it can be applied to ordinal or survival outcome data as well as more complicated association tests used in next-generation sequencing studies. We show that our method results in increased power for detecting genetic risk factors for idiopathic pain conditions in data collected from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study.

email: ebair@email.unc.edu

5. POSTERS: SURVIVAL ANALYSIS

5a. CENSORED QUANTILE REGRESSION WITH RECURSIVE PARTITIONING BASED WEIGHTS

Andrew Wey*, University of Minnesota
Lan Wang, University of Minnesota
Kyle Rudser, University of Minnesota

Censored quantile regression provides a useful alternative to the Cox model for analyzing survival data. It directly models the conditional quantile of the survival time, hence is easy to interpret. Moreover, it relaxes the proportionality constraint on the hazard function and allows for modeling heterogeneity of the data. Recently, Wang and Wang (2009) proposed a locally weighted censored quantile regression approach which allows for covariate-dependent censoring and is less restrictive than other censored quantile regression methods. However, their kernel smoothing based weighting scheme requires all covariates to be continuous and encounters practical difficulty with even a moderate number of covariates. We propose a new weighting approach that uses recursive partitioning, e.g., survival trees, that offers greater flexibility in handling covariate-dependent censoring in moderately high dimension and can be applied to both continuous and discrete covariates. We prove that this new weighting scheme leads to consistent estimation of the quantile regression coefficients and demonstrate the effectiveness of the new approach via Monte Carlo simulations. We illustrate the new method using a data set from a clinical trial on primary biliary cirrhosis.

email: weyxx003@morris.umn.edu

5b. NONPARAMETRIC COMPARISON OF SURVIVAL FUNCTIONS BASED ON INTERVAL CENSORED DATA WITH UNEQUAL CENSORING

Ran Duan*, University of Missouri, Columbia
Yanqing Feng, Wuhan University
Tony (Jianguo) Sun, University of Missouri, Columbia

Nonparametric comparison of survival functions is one of the most commonly required task in failure time studies such as clinical trials and for this, many procedures have been developed under various situations (Kalbfleisch and Prentice, 2002; Sun, 2006). This paper considers a situation that often occurs in practice but has not been discussed much: the comparison based on interval-censored data in the presence of unequal censoring. That is, one observes only interval-censored data and the distributions of or the mechanisms behind censoring variables may depend on treatments and thus be different for the subjects in different treatment groups. For the problem, a test procedure is developed that takes into account the difference between the distributions of the censoring variables, and the asymptotic normality of the test statistics is given. For the assessment of the performance of the procedure, a simulation study is conducted and suggests that it works well for practical situations. An illustrative example is provided.

email: duanran25@gmail.com

5c. SMALL SAMPLE PROPERTIES OF LOGRANK TEST WITH HIGH CENSORING RATE

Yu Deng*, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill

Logrank test is commonly used for comparing survival distributions between treatment and control groups. When censoring rate is low and the sample size is moderate, the approximation based on the asymptotic normal distribution of the logrank test works well in finite samples. However, in some studies, the sample size is small (e.g. 10, 20 per group) and the censoring rate is high (e.g. 0.8, 0.9). Under such situation, we conduct a series of simulations to compare the performance of the logrank test based on normal approximation, permutation, and bootstrap. In general, the type I error rate based on the bootstrap test is slightly inflated, while the permutation and normal approximation are relatively conservative. The bootstrap test has a higher power among all the three tests when each group has more than one failure. In addition, when each group has only one failure, the power based on the permutation test is slightly higher than the other two tests. In conclusion, when the hazard ratio is larger than 2.0 and the number of failure is ranging from 2 to 20 for each group, the bootstrap test is more powerful than the logrank test.

email: yudeng@live.unc.edu

5d. STRATIFIED AND UNSTRATIFIED LOG-RANK TEST IN CORRELATED SURVIVAL DATA

Yu Han*, University of Rochester
David Oakes, University of Rochester
Changyong Feng, University of Rochester

The log-rank test is the most widely used nonparametric method for testing treatment differences in survival analysis due to its efficiency under the proportional hazards model. Most previous work on the log-rank test has assumed that the samples from different treatment groups are independent. This assumption is not always true. In multi-center clinical trials, survival times of patients in the same medical center may be correlated due to factors specific to each center. For such data we can construct both stratified and unstratified log-rank tests. These two tests turn out to have very different powers for correlated samples. An appropriate linear combination of these two tests may give a more powerful test than either individual test. Under a frailty model, we obtain a closed form of asymptotic local alternative distributions and the correlation coefficient between these two tests. Based on these results we construct an optimal linear combination of the two test statistics to maximize the local power. We also consider sample size calculation in the paper. Simulation studies are used to illustrate the robustness of the combined test.

email: Yu_Han@urmc.rochester.edu

5e. ANALYSIS OF MULTIPLE MYELOMA LIFE EXPECTANCY USING COPULA

Eun-Joo Lee*, Millikin University

Multiple myeloma is a blood cancer that develops in the bone marrow. It is assumed that in most cases multiple myeloma develops in association with several medical factors acting together, although the leading cause of the disease has not yet been identified. In this paper, we investigate the relationship between the factors to measure multiple myeloma patients' survival time. For this, we employ a copula that provides a convenient way to construct statistical models for multivariate dependence. Through an approach via copulas, we find the most influential medical factors that affect the survival time. Some goodness-of-fit tests are also performed to check the adequacy of the copula chosen for the best combination of the survival time and the medical factors. Using the Monte Carlo simulation technique with the copula, we re-sample survival times from which the anticipated life span of a patient with the disease is calculated.

email: elee@millikin.edu

5f. FRAILTY PROBIT MODEL FOR CLUSTERED INTERVAL-CENSORED FAILURE TIME DATA

Haifeng Wu*, University of South Carolina, Columbia
Lianming Wang, University of South Carolina, Columbia

Clustered interval-censored data commonly arise in many studies of biomedical research where the failure time of interest is subject to interval-censoring and subjects are correlated for being in the same cluster. In this paper, we propose a new frailty semiparametric Probit regression model to study the covariate effects on the failure time and the intra-cluster dependence. The proposed normal frailty Probit model enjoys several nice properties that existing survival models do not have: (1) the marginal distribution of the failure time is a semiparametric Probit model, (2) the regression parameters can be interpreted either as the conditional covariate effects given frailty or the marginal covariate effects up to a multiplicative constant, and (3) the intra-cluster association can be summarized by two nonparametric measures in simple and closed form. A fully Bayesian estimation method is developed based on the use of monotone spline for the unknown nondecreasing function and our Gibbs sampler is straightforward to implement. The proposed method performs well in estimating the regression parameters and is robust to misspecified frailty distributions in our simulation studies. Two real-life data sets are analyzed as illustrations.

email: wuh@email.sc.edu

5g. SEMIPARAMETRIC ACCELERATE FAILURE TIME MODELING FOR CLUSTERED FAILURE TIMES FROM STRATIFIED SAMPLING

Sy Han Chiou*, University of Connecticut
Sangwook Kang, University of Connecticut
Jun Yan, University of Connecticut

Clustered failure time data arising from stratified sampling are often encountered in studies where it is desired to reduce both cost and sampling error. In such settings, semiparametric accelerated failure time (AFT) models have not been used as frequently as Cox relative risk models in practice due to lack of efficient and reliable computing routines for statistical inferences, with challenge rooted in nonsmooth rank-based estimating functions. The recently proposed induced smoothing approach, which provides fast and accurate inferences for AFT models, is generalized to incorporate weights that can be applied to accommodate stratified sampling design. The estimators resulting from the induced smoothing weighted estimating equations are consistent and asymptotically normal with the same distribution as the estimators from the nonsmooth estimating equations. The variance is estimated by two computationally efficient sandwich estimators. The proposed method is validated in extensive simulation studies and appears to provide valid inferences. In a stratified case-cohort design with clustered times to tooth extraction in a dental study, similar results to those from alternative methods were found but were obtained much faster.

email: steven.chiou@uconn.edu

5h. A CUMULATIVE INCIDENCE JOINT MODEL OF TIME TO DIALYSIS INDEPENDENCE AND INFLAMMATORY MARKER PROFILES IN ACUTE KIDNEY INJURY

Francis Pike*, University of Pittsburgh
Jonathan Yabes, University of Pittsburgh
John Kellum, University of Pittsburgh

Recovery from Acute Renal Failure is a clinically relevant issue in critical care medicine. The central goal of the Biological Markers of Recovery for the Kidney study (BIOMARK) was to understand the relationship between inflammation and oxidative stress in recovery from Acute Renal Failure (ARF) and how intensity of Renal Replacement Therapy (RRT) affects this relationship. To effectively model this relationship the chosen analytical procedure has to, (i) account for censoring in patient inflammatory profiles due to the sensitivity of the assays, and (ii) be able to include this information into the survival model whilst accounting for competing terminal events such as death. To this end we formulated and implemented a fully parametric cumulative incidence (CIF) joint model within SAS using NLMIXED. Specifically we combined a linear mixed effects Tobit model with a parametric CIF model proposed by Jeong and Fine to account for the longitudinal censoring and competing risks respectively. We verified the performance of this model via simulation and applied this method to the BIOMARK study to ascertain if intensity of treatment and inflammatory profiles significantly affect time to dialysis independence.

email: pikef@upmc.edu

5i. AGE-SPECIFIC RISK PREDICTION WITH LONGITUDINAL AND SURVIVAL DATA

Wei Dai*, Harvard School of Public Health
Tianxi Cai, Harvard School of Public Health
Michelle Zhou, Simon Fraser University

Often in cohort studies where the primary endpoint is time to an event, patients are also monitored longitudinally with respect to one or more biological variables throughout the follow-up period. A primary goal of such studies is to predict the risk of future events. Joint models for both the longitudinal process and survival data have been developed in recent years to analyze such data. In the joint modeling framework, risk of future events is estimated using Monte Carlo simulations. In most existing risk prediction models, age is modeled as one of the standard risk factors with simple effects. However, for many complex phenotypes such as the cardiovascular

disease, important risk factors such as BMI might have substantially different effect on future risks depending on the age. Hence the longitudinally collected risk factor information on the same patient might contribute information to different age specific risks depending on when the risk factors are measured. To incorporate such age varying effects, we introduce an alternative approach that estimates the age-specific risk directly via a flexible varying-coefficient model. The performance of our method is explored using simulation studies. We illustrate this method using the Framingham Heart Study data and compare our prediction with the Framingham score.

email: wdai.0102@gmail.com

5j. A FRAILTY-BASED PROGRESSIVE MULTISTATE MODEL FOR PROGRESSION AND DEATH IN CANCER STUDIES

Chen Hu*, Radiation Therapy Oncology Group/American College of Radiology
Alex Tsodikov, University of Michigan

In advanced or adjuvant cancer studies, progression-related events (e.g., progression-free or recurrence-free survival) and cancer death are common endpoints that are sequentially observed. The relationship between covariate (e.g., therapeutic intervention), progression, and death is often of interest, as it may provide a key to optimal treatment decisions. The evaluation of this relationship is often complicated by the latency of disease progression leading to undetected or missing progression-related events. We consider a progressive multistate model with a frailty modeling the association between progression and death, and propose a semiparametric regression model for the joint distribution. An Expectation-Maximization (EM) approach is used to derive the maximum likelihood estimators of covariate effects on both endpoints, the probability of missing progression event, as well as the parameters involved in the association. The asymptotic properties of the estimators are studied. We illustrate the proposed method with Monte Carlo simulation and data analysis of a clinical trial of colorectal cancer adjuvant therapy.

email: chenhu@umich.edu

5k. TIME-DEPENDENT ROC ANALYSIS USING DATA WITH OUTCOME-DEPENDENT SAMPLING BIAS

Shanshan Li*, Johns Hopkins School of Public Health
Mei-Cheng Wang, Johns Hopkins School of Public Health

This paper considers estimation of time-dependent receiver operating characteristic (ROC) curves when survival data are collected subject to sampling bias. The sampling bias exists in many follow-up studies where data are observed according to a cross-sectional sampling scheme which tends to over sample individuals with longer survival times. To correct the sampling bias, we develop a semiparametric estimation method for estimating time-dependent ROC curves under proportional hazards assumption. The proposed estimators are consistent and converge to Gaussian processes, while substantial bias may arise if standard estimators for right-censored data are used. Statistical inference is also established to identify the optimal combination of multiple markers under the proportional hazards model. To illustrate our method, we analyze data from an Alzheimer's disease study and estimate ROC curves that assess how well the cognitive measurements can distinguish subjects that progress to mild cognitive impairment from subjects that remain normal.

email: shli@jhsph.edu

5l. IMPUTATION GOODNESS-OF-FIT TESTS FOR LENGTH-BIASED AND RIGHT-CENSORED DATA

Na Hu*, University of Missouri, Columbia
Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Jianguo Sun, University of Missouri, Columbia

In recent years, there has been a rising interest in length-biased survival data, which are commonly encountered in epidemiological cohort studies, cancer screening trials and many others. This paper considers checking the adequacy of a parametric distribution with length-biased data. We propose a new one-sample Kolomogorov-Smirnov type of goodness-of-fit test based on the imputation idea. Its large sample properties can be easily derived. Simulation studies are conducted to assess the performance of the test and compare it with the existing test. Finally, we use the data from a prevalent cohort study of patients with dementia to illustrate the proposed methodology.

email: nh2hd@mail.missouri.edu

5m. A WEIGHTED APPROACH TO ESTIMATION IN AFT MODEL FOR RIGHT-CENSORED LENGTH-BIASED DATA

Chetachi A. Emeremni*, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

When length-biased data are subjected to right censoring, analysis of such data could be quite challenging because of the induced informative censoring, since survival time and censoring time are correlated through a common backward event initiation time. We propose a weighted estimating equation approach to estimate the parameters of a length-biased regression model, where the residual censoring time is assumed to depend on the truncation times through a parametric model, and the estimating equation is weighted for both length bias and right censoring. We establish the asymptotic properties of our estimators. Simulation results show that the proposed estimator is more efficient than existing methods and is easier to apply. We apply our methods to the Channing House data.

email: cemeremn@uthsc.edu

5n. SEMIPARAMETRIC APPROACH FOR REGRESSION WITH COVARIATE SUBJECT TO LIMIT OF DETECTION

Shengchun Kong*, University of Michigan
Bin Nan, University of Michigan

We consider regression analysis with left-censored covariate due to the lower limit of detection (LOD). The complete case analysis by eliminating observations with values below LOD may yield valid estimates for regression coefficients, but is less efficient. Existing substitution or maximum likelihood method usually relies on parametric models for the unobservable tail probability, thus may suffer from model misspecification. To obtain robust results, we propose a likelihood based approach for the regression parameters using a semiparametric accelerated failure time model for the covariate that is left-censored by LOD. A two-stage estimation procedure is considered, where the conditional distribution of the covariate with LOD given other variables is estimated first before maximizing the likelihood function for the regression parameters. The proposed method outperforms the traditional complete case analysis and the simple substitution methods in simulation studies. Technical conditions for desirable asymptotic properties will be discussed.

email: kongsc@umich.edu

5o. A WEIGHTED ESTIMATOR OF ACCELERATED FAILURE TIME MODEL UNDER PRESENCE OF DEPENDENT CENSORING

Youngjoo Cho*, The Pennsylvania State University
Debashis Ghosh, The Pennsylvania State University

Independent censoring is one of the crucial assumptions in models of survival analysis. However, this is impractical in many medical studies, where the presence of dependent censoring leads to difficulty in analyzing covariate effects on disease outcomes. The semicompeting risks framework proposed by Lin et al. (1996, Biometrika) and Peng and Fine (2006, Journal of the American Statistical Association) is a suitable approach to handling dependent censoring. These authors proposed estimators based on an artificial censoring technique. However, they did not consider efficiency of their estimators in detail. In this paper, we propose a new weighted estimator for the accelerated failure time (AFT) model under dependent censoring. One of the advantages in our approach is that these weights are optimal among all the linear combina-

tions of these two estimators previously referenced. Moreover, to calculate these weights, a novel resampling-based scheme is employed. Attendant asymptotic statistical results for the estimator are established. In addition, simulation studies, as well as application to real data, show the gains in efficiency for our estimator.

email: yvc5154@psu.edu

5p. NON-PARAMETRIC CONFIDENCE BANDS FOR SURVIVAL FUNCTION USING MARTINGALE METHOD

Seung-Hwan Lee*, Illinois Wesleyan University

A simple computer-assisted method of constructing non-parametric simultaneous confidence bands for survival function with right censored data is introduced. This method requires no distributional assumptions. The procedures are based on the integrated martingale process whose distribution is approximated by a Gaussian process. The supremum distribution of the Gaussian process generated by simulation leads to a construction of the confidence bands. To improve the inference procedures for the finite sample sizes, the log-minus-log transformation is employed. The newly developed procedures for estimating the confidence bands are assessed through simulation and applied to a real-world data set regarding leukemia.

email: slee2@iwu.edu

6. POSTERS: LONGITUDINAL AND MISSING DATA

6a. POOLED CORRELATION COEFFICIENTS FOR LONGITUDINALLY MEASURED BIOMARKERS

Su Chen*, University of Michigan
Thomas M. Braun, University of Michigan

We wish to determine if the pair-wise correlations of time-adjacent longitudinal data are homogeneous and can be pooled into a single time-invariant value. After testing the homogeneity hypothesis, and lack of significance is found, a decision can be made to pool the various correlation coefficients into a common correlation between two measured biomarkers. We propose an estimator of the pooled correlation coefficient based on Mantel-Haenszel methods and derive a variance estimate of this estimator. The bias of our estimator and its variance estimate is evaluated in different settings via simulation.

email: such@umich.edu

6b. LONGITUDINAL ANALYSIS OF THE EFFECT OF HEALTH TRAITS ON RELATIONSHIPS IN A SOCIAL NETWORK

A James O'Malley*, Harvard Medical School
Sudeshna Paul, Emory University

We develop a new longitudinal model for relationship status of pairs of individuals ('dyads') in a social network. We first consider the relationship status of a single dyad, which in the case of binary relationships follows a four-component multinomial distribution. To extend the model to the whole network we account for the dependence of observations between dyads by assuming dyads are conditionally independent given actor-specific latent variables and lagged covariates judiciously chosen to account for important inter-dyad dependencies (e.g., transitivity "a friend of a friend is a friend"). Model parameters are estimated using Bayesian analysis implemented via Markov chain Monte Carlo (MCMC). The model is illustrated using a friendship network constructed from a panel study on the health and lifestyles of teenaged girls attending a school in Scotland. Results of the analysis indicate a strong dependence across time, high reciprocation of ties between individuals, extensive triadic clustering, but did not find strong evidence of homophily (the tendency towards ties forming and continuing between individuals with similar traits). Examination of model fit revealed that our model successfully captured the most important features of the data.

email: omalley@hcp.med.harvard.edu

6c. A MARKOV TRANSITION MODEL FOR LONGITUDINAL ORDINAL DATA WITH APPLICATIONS TO KNEE OSTEOARTHRITIS AND PHYSICAL FUNCTION DATA

Huiyong Zheng*, University of Michigan
Carrie Karvonen-Gutierrez, University of Michigan
Siobàn D. Harlow, University of Michigan

In women, the menopausal transition occurs over several years. Marked physiological changes occur across this transition including initiation and progression of knee osteoarthritis (OAK) and declines in physical functioning, where the measures are categorized into ordinal outcomes or states. Understanding the bidirectional development of OAK disease status and improvement or deterioration of physical function is of great public health and clinical interest. In longitudinal studies with ordinal outcomes, each individual experiences a finite number of states and may transit from state to state across time. Quantification of these dynamics over time requires assessment of a multistate stochastic transition process. We propose a mixed-effect Markov transition model to analyze such stochastic process with finite state space. Model parameters are estimated using maximum-likelihood estimation. The homogeneity of dynamic transition behavior with time-dependent covariates (risk factors) will be evaluated. We illustrate the method by modeling two longitudinal ordinal outcomes, 15 years (6 follow-up visits) of OAK data scored by the Kellgren and Lawrence system (0=normal/no disease, 1=doubtful OA, 2=minimal OA, 3=moderate OA, and 4=severe OA), and 10 years (5 follow-up visits) of self-reported physical functioning limitations (SF-36) with classifications of 0=not limited at all, 1=a little limited, or 2=limited a lot.

email: zhenghy@umich.edu

6d. ZERO-INFLATION IN CLUSTERED BINARY RESPONSE DATA: MIXED MODEL AND ESTIMATING EQUATION APPROACHES

Kara A. Fulton*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Danping Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Paul S. Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

The NEXT Generation Health study investigates the dating violence of 2776 tenth-grade students using a survey questionnaire. Each student is asked to affirm or deny multiple instances of violence in his/her dating relationship. There is, however, evidence suggesting that students not in a relationship responded to the survey, resulting in both structural and sampling zeros. This paper proposes likelihood-based and estimating equation approaches to analyze the zero-inflated clustered binary response data. We adopt a mixed model method to account for the cluster effect, and the model parameters are estimated using a maximum likelihood approach that requires a Gaussian-Hermite quadrature (GHQ) approximation for implementation. Since an incorrect assumption on the random effects distribution may bias the results, we construct generalized estimating equations (GEE) that do not require the correct specification of within-cluster correlation. In a series of simulation studies, we examine the performance of maximum-likelihood and GEE in terms of their small sample bias, efficiency and robustness. We illustrate the importance of properly accounting for this zero inflation by re-analyzing NEXT data where this issue has previously been ignored.

email: fultonka@mail.nih.gov

6e. THE USE OF TIGHT CLUSTERING TECHNIQUE IN DETERMINING LATENT LONGITUDINAL TRAJECTORY GROUPS

Ching-Wen Lee*, University of Pittsburgh
Lisa A. Weissfeld, University of Pittsburgh

Latent group-based modeling is widely used to categorize individuals into several homogeneous trajectory groups. With latent group-based modeling, all individuals are forced into a specific group either forcing a larger number of groups to be formed or forcing small groups of individuals into one of the large groups when only two or three groups are specified. In the former case when groups with a small number of individuals are identified, analyses using group membership as a covariate may be unstable. To address these issues, we propose to use the tight clustering technique (Tseng and Wong, 2005) to identify prominent latent trajectory groups and to differentiate them from a miscellaneous group which will contain individuals whose trajectory patterns are dissimilar to the patterns in the rest of the population. An individual will be assigned to a specific trajectory group if their corresponding posterior probability of belonging to that group is greater than or equal to a predetermined cutoff value (e.g., 0.8). We use the Bayesian information criterion as the criterion for model selection. The performance of the proposed method is evaluated through a simulation study and a clinical example is given for demonstration purposes.

email: chl98@pitt.edu

6f. A NOVEL SEMI-PARAMETRIC APPROACH FOR IMPUTING MIXED DATA

Irene B. Helenowski*, Northwestern University
Hakan Demirtas, University of Illinois at Chicago

Multiple imputation via joint modeling is a popular approach to handling missing mixed data which includes continuous and binary or categorical variables. Joint modeling used in imputing mixed data involves the general location model. But what if assumptions of this model are violated? We thus propose a semi-parametric approach for imputing mixed data which allows us the relax assumptions of the general location model. Simulation studies and real data applications indicate promising results.

email: i-helenowski@northwestern.edu

6g. HOT DECK IMPUTATION OF NONIGNORABLE MISSING DATA WITH SENSITIVITY ANALYSIS

Danielle M. Sullivan*, The Ohio State University
Rebecca Andridge, The Ohio State University

Hot deck imputation is a common method for handling item nonresponse in surveys, but most implementations assume data are missing at random. Here, we combine the Approximate Bayesian Bootstrap (ABB) distance-based donor selection method of Siddique and Belin (2008) with the Proxy Pattern-Mixture (PPM) model (Andridge and Little 2011). The proxy pattern-mixture model is used to define distances between donors and donees under different assumptions on the missingness mechanism. This creates a proxy hot deck with distance-based donor selection to perform imputation, accompanied by an intuitive sensitivity analysis. As with the parametric PPM model, missingness in the outcome is assumed to be a linear function of the outcome and the proxy variable, estimated from a regression analysis of respondent data. The sensitivity analysis allows for simple comparisons between ignorable and varying levels of nonignorable missingness. The PPM hot deck provides a more concise sensitivity analysis than using the more than 6 various 'shaped' ABBs of Siddique and Belin. Compared to the parametric PPM model, the PPM hot deck is potentially less sensitive to model misspecification. We compare the bias and coverage of estimates from the PPM hot deck with the ABB hot deck through simulations and apply the method to data from the Ohio Family Health Survey.

email: sullivan.467@osu.edu

6h. IMPUTING MISSING CHILD WEIGHTS IN GROWTH CURVE ANALYSES

Paul Kolm*, Christiana Care Health System
Deborah Ehrenthal, Christiana Care Health System
Matthew Goldshore, John Hopkins University

Missing data present a challenge for analysis with respect to results of the analysis and conclusions made on the basis of the results. A complete case (CC) analysis or filling in missing values with averages has been shown to result in erroneous conclusions had complete data been available. More sophisticated methods of imputing missing values have been developed for data that are missing at random (MAR) and not missing at random (NMAR). The purpose of this study was to investigate whether differences in methods of imputing missing weight values affect the estimation of children's growth curves. Longitudinal data from 4,000+ mother-child dyads were obtained to explore the association of women's characteristics and risk factors during gestation with the development of overweight/obesity in their offspring at age 4. As would be expected, there were a significant number of missing weights over a 4-year period. We compare estimates for last-value-carried forward (LVCF), simple linear interpolation of individual child weights, multiple imputation assuming MAR and multiple imputation assuming NMAR.

email: pkolm@christianacare.org

6i. STUDY OF SEXUAL PARTNER ACCRUAL PATTERNS AMONG ADOLESCENT WOMEN VIA GENERALIZED ADDITIVE MIXED MODELS

Fei He*, Indiana University School of Medicine
Jaroslaw Harezlak, Indiana University Schools of Public Health and Medicine
Dennis J. Fortenberry, Indiana University School of Medicine

The number of lifetime partners is a consistently identified epidemiological risk factor for sexually transmitted infections (STIs). Higher rate of partner accrual during adolescence has been associated with increased STI rates among adolescent women. To study sexual partner accrual pattern among adolescent females, we applied generalized additive mixed models (GAMM) to the data obtained from a longitudinal STI study. GAMM regression components included a bivariate function enabling separation of cohort ('age at study entry') and longitudinal ('follow-up years') effects on partner accrual while the correlation was accounted for by the subject-specific random components. Longitudinal effect partial derivative was used to estimate within-subject rates of partner accrual and their standard errors. The results show that slowing of partner accrual depends more on the prior sexual experience and less on the females' chronological age. Our modeling approach combining the GAMM flexibility and the time covariates' of interest definition enabled clear differentiation between the cohort (chronological age) and longitudinal (follow-up time) effects, thus providing the estimates of both between-subject differences and within-subject trajectories of partner accrual.

email: hefei@iupui.edu

6j. STATISTICAL ANALYSIS WITH MISSING EXPOSURE DATA MEASURED BY PROXY RESPONDENTS: A MISCLASSIFICATION PROBLEM EMBEDDED IN A MISSING-DATA PROBLEM

Michelle Shardell*, University of Maryland School of Medicine

Researchers often recruit proxy respondents, such as relatives or caregivers, for studies of older adults when study participants cannot provide self-reports (e.g., due to illness). Proxies are often only recruited to report on participants with missing self-reports; thus, either a proxy report or participant self-report, but not both, is available for each participant. When exposures are binary and participant self-reports are the gold standard, substituting proxy reports for missing participant self-reports can introduce misclassification error and produce biased estimates. Also, the missing-data mechanism for

participant self-reports is not identifiable and may be informative. Most methods with a misclassified exposure require either validation data, replicate data, or an assumption of nondifferential misclassification. We propose a pattern-mixture model where none of these is required. Instead, the model is indexed by two user-specified tuning parameters that represent an assumed level of agreement between the observed proxy and missing participant responses and can be varied to perform a sensitivity analysis. We estimate associations standardized for high-dimensional covariates using multiple imputation followed by inverse probability weighting. Simulation studies show that the proposed method performs well.

email: mshardel@epi.umaryland.edu

7. POSTERS: IMAGING / HIGH DIMENSIONAL DATA

7a. HOMOTOPIC GROUP ICA

Juemin Yang*, Johns Hopkins University
Ani Eloyan, Johns Hopkins University
Anita Barber, Kennedy Krieger Institute
Mary Beth Nebel, Kennedy Krieger Institute
Stewart Mostofsky, Kennedy Krieger Institute
James Pekar, Kennedy Krieger Institute
Brian Caffo, Johns Hopkins University

Independent Component Analysis (ICA) is a computational technique for revealing hidden factors that underlie sets of measurements or signals. It is widely used in a variety of academic fields such as functional neuroimaging. We have devised a new group ICA approach, Homotopic group ICA (H-gICA), for blind source separation of fMRI data. Our new approach enables us to double the sample size via brain functional homotopy, the high degree of synchrony in spontaneous activity between geometrically corresponding interhemispheric regions. H-gICA can increase the power for finding underlying networks with the existence of noise and it is proved theoretically to be the same as commonly used group ICA when the true sources are perfectly homotopic and noise free. Moreover, compared to commonly used ICA algorithms, the structure of the H-gICA input data leads to significant improvement of computational efficiency. In a simulation study, our approach proved to be effective in both homotopic and non-homotopic settings. We also show the effectiveness of our approach by its application on the ADHD-200 dataset. Out of the 15 components postulated by H-gICA, several brain networks were found including: the visual network, the default mode network, the auditory network, etc. In addition to improving network estimation, H-gICA allows for the investigation of functional homotopy via ICA based networks.

email: juyang@jhsph.edu

7b. THE 'GENERAL LINEAR MODEL' IN fMRI ANALYSIS

Wenzhu Mowrey*, Albert Einstein College of Medicine

Functional Magnetic Resonance Imaging (fMRI) has seen wide use in neuropsychology since its invention in the early 1990s. Statisticians have made critical contributions along its side. The interdisciplinary nature of an imaging study makes the learning curve steep for everybody because of the physics, the neuroanatomy and the various imaging processing involved. This presentation is to introduce fMRI data analysis stream surrounding its core - the general linear model (LM). We will focus on the commonly used task data, where subjects perform tasks, for example finger tapping, in a scanner and the goal is to find the task-related brain activities and to assess its differences across groups. We start with preprocessing steps including slice timing correction, motion correction, normalization and smoothing. Then we perform the subject level (first level) general linear model analysis, where effects of interest are extracted from each individual's time series data. We will clarify its differences from the LM in a traditional statistics setting and the role of the hemodynamics response function in the model. The last step is the group level (second level) analysis, where a simple review on multiple comparison correction will be given.

email: wenzhu.mowrey@einstein.yu.edu

7c. OASIS IS AUTOMATED STATISTICAL INFERENCE FOR SEGMENTATION WITH APPLICATIONS TO MULTIPLE SCLEROSIS LESION SEGMENTATION IN MRI

Elizabeth M. Sweeney*, Johns Hopkins University
Russell T. Shinohara, University of Pennsylvania
Navid Shiee, Henry M. Jackson Foundation
Farrah J. Mateen, Johns Hopkins University
Avni A. Chudgar, Brigham and Women's Hospital and Harvard Medical School
Jennifer L. Cuzzocreo, Johns Hopkins University
Peter A. Calabresi, Johns Hopkins University
Dzung L. Pham, Henry M. Jackson Foundation
Daniel S. Reich, National Institute of Neurological Disease and Stroke, National Institutes of Health
Ciprian M. Crainiceanu, Johns Hopkins University

We propose OASIS is Automated Statistical Inference for Segmentation (OASIS), an automated statistical method for segmenting multiple sclerosis (MS) lesions in magnetic resonance images (MRI). We use logistic regression models incorporating multiple MRI modalities to estimate voxel-level probabilities of lesion presence. Intensity-normalized T1-weighted, T2-weighted, fluid-attenuated inversion recovery and proton density volumes from 131 MRI studies with manual lesion segmentations are used to train and validate our model. Within this set, OASIS detected lesions with an area under the receiver-operator characteristic curve of 98% (95% CI; [96%, 99%]) at the voxel level. Use of intensity-normalized MRI volumes enables OASIS to be robust to variations in scanners and acquisition sequences. We applied OASIS to 169 MRI studies acquired at a separate imaging center. A neuroradiologist compared these segmentations to segmentations produced by another software, LesionTOADS. For lesions, OASIS out-performed LesionTOADS in 77% (95% CI: [71%, 83%]) of cases. For a randomly selected subset of 50 of these studies, one additional radiologist and one neurologist also scored the images. Within this set, the neuroradiologist ranked OASIS higher than LesionTOADS in 76% (95% CI: [64%, 88%]) of cases, the neurologist 66% (95% CI: [52%, 78%]) and the radiologist 52% (95% CI: [38%, 66%]).

email: emsweene@jhsph.edu

7d. NONLINEAR MIXED EFFECTS MODELING WITH DIFFUSION TENSOR IMAGING DATA

Namhee Kim*, Albert Einstein College of Medicine of Yeshiva University
Craig A. Branch, Albert Einstein College of Medicine of Yeshiva University
Michael L. Lipton, Albert Einstein College of Medicine of Yeshiva University

In many statistical analyses with imaging data, a summary value approach for each region, e.g. an average of observed brain physiology across voxels, has been frequently adopted. However, these approaches ignore spatial variability within each region, and have potential biases in the estimated parameters. We thus need approaches incorporating individual voxels to reduce biases and enhance efficiency in estimation of parameters. Fractional Anisotropy (FA) from diffusion tensor imaging (DTI) describes the degree of anisotropy of a diffusion process of water molecules, and abnormally low FA has been consistently found with traumatic brain injury (TBI). Since soccer heading may form a repetitive mild TBI, we in this study investigate the trajectory of FA over soccer heading exposures by employing a growth curve. Proposed nonlinear mixed effects (NLME) model includes random effects to account spatial variability of each parameter of the growth curve across voxels, and adopt simultaneous autoregressive model to account correlation among neighboring voxels. The fitted NLME model was compared to a nonlinear model which utilizes an average of FA values per subject. The proposed NLME model is more efficient and provides additional information on spatial variability of the estimated parameters of the growth curve.

email: namhee.kim@einstein.yu.edu

7e. CLUSTERING OF HIGH-DIMENSIONAL LONGITUDINAL DATA

Seonjoo Lee*, Henry Jackson Foundation
Vadim Zipunnikov, Johns Hopkins University
Brian S. Caffo, Johns Hopkins University
Ciprian Crainiceanu, Johns Hopkins University
Dzung L. Pham, Henry Jackson Foundation

We focus on uncovering latent classes of subjects with sparsely observed high-dimensional longitudinal data. Such situations commonly occur in longitudinal structural brain image analysis. The current existing clustering algorithms are not directly applicable nor do those algorithms exploit longitudinal information. We propose a distance metric between two high-dimensional data trajectories in the presence of substantial visit-to-visit errors. Longitudinal functional principal component analysis (LFPCA) provides a framework to reduce the dimensionality of data onto a lower-dimensional, subject-specific subspace. We show that the distance between two high-dimensional longitudinal data trajectories is equivalent to the subject-specific LFPCA scores. Our simulation studies evaluate the performance of various clustering analysis based on the proposed distance and other dimension reduction methods. Application of the approach to longitudinal brain images acquired from a multiple sclerosis population reveals two distinct clusters and the pattern is found to be associated with clinical scores.

email: seonjool@gmail.com

7f. MULTIPLE COMPARISON PROCEDURES FOR NEUROIMAGING GENOMEWIDE ASSOCIATION STUDIES

Wen-Yu Hua*, The Pennsylvania State University
Thomas E. Nichols, University of Warwick, U.K.
Debashis Ghosh, The Pennsylvania State University

Recent research in neuroimaging has been focusing on assessing associations between genetic variants measured on a genomewide scale and brain imaging phenotypes. Many publications in the area use massively univariate analyses on a genomewide basis for finding single nucleotide polymorphisms that influence brain structure. In this work, we propose using various dimensionality reduction methods on both brain MRI scans and genomic data, motivated by the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. We also consider a new multiple testing adjustments inspired from the idea of local false discovery rate of Efron et al. (2001). Our proposed procedure is able to find associations between genes and brain regions at a better significance level than in the initial analyses.

email: wxh182@psu.edu

7g. TESTS OF THE MONOTONICITY AND CONVEXITY IN THE PRESENCE OF CORRELATION AND THEIR APPLICATION ON DESCRIBING MOLECULE STRUCTURE

Huan Wang*, Colorado State University
Mary C. Meyer, Colorado State University
Jean D. Opsomer, Colorado State University
F. J. Breidt, Colorado State University

Methods for testing the shape of a function, such as monotonicity and/or convexity, are useful in many applications, especially for time series data. In this talk, we propose a set of hypothesis tests using regression splines and shape restricted inference in the presence of stationary autocorrelated errors. The null hypothesis H_{0} is that the function is constant/linear, H_{1} is that the function is constrained to be monotone/convex, and H_{2} is that the function is unconstrained. The likelihood ratio test statistic of H_{0} versus H_{1} has exact null distribution if the covariance matrix of errors is known and has nice asymptotic behavior if the covariance matrix is unknown. The test of H_{1} versus H_{2} uses an estimate of the distribution of the minimum slope/second derivative of the spline estimator under the null hypothesis and proved to behave nicely both for small sample size and asymptotically. The test that H_{1}: the function is decreasing and convex, versus H_{2}: the function is unconstrained, is applied to intensity data from small angle X-ray scattering (SAXS) experiments. The proposed method serves as a useful pretest in this context, because under H_{1}, a classical regression-based procedure for estimating a molecule's radius of gyration can be applied.

email: wangh@stat.colostate.edu

7h. STRUCTURED FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

Haochang Shou*, Johns Hopkins Bloomberg School of Public Health
Vadim Zipunnikov, Johns Hopkins Bloomberg School of Public Health
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health
Sonja Greven, Ludwig-Maximilians-Universitat, Germany

Motivated by modern observational studies, we introduce a wide class of functional models that expands classical nested and crossed designs. Our approach targets a better interpretability of level-specific features through natural inheritance of designs and explicit modeling of correlations. We base our inference on functional quadratics and their relationship with underlying covariance structure of the latent processes. For modeling populations of ultra-high dimensional functions or images with known structure induced by sampling or experimental design, we develop a computationally fast and scalable estimation procedure. We illustrate the methods with three data sets that represent a new generation of functional observation: a high-frequency accelerometer data collected for estimating daily energy expenditure, pitch linguistic data used for phonetic analysis, and EEG data representing electrical brain activity during sleep.

email: haochang.shou@gmail.com

7i. FUNCTIONAL DATA ANALYSIS OF TREE DATA OBJECTS

Dan Shen*, University of North Carolina, Chapel Hill
Haipeng Shen, University of North Carolina, Chapel Hill
Shankar Bhamidi, University of North Carolina, Chapel Hill
Yolanda Muñoz Maldonado, Research Institute, Beaumont Health System
Yongdai Kim, Seoul National University
Steve Marron, University of North Carolina, Chapel Hill

Data analysis on non-Euclidean spaces, such as tree spaces, can be challenging. The main contribution of this paper is establishment of a connection between tree data spaces and the well developed area of Functional Data Analysis (FDA), where the data objects are curves. This connection comes through two tree representation approaches, the Dyck path representation and the branch length representation. These representations of trees in Euclidean spaces enable us to exploit the power of FDA to explore statistical properties of tree data objects. A major challenge in the analysis is the sparsity of tree branches in a sample of trees. We overcome this issue by using a tree pruning technique that focuses the analysis on important underlying population structures. This method parallels scale-space analysis in the sense that it reveals statistical properties of tree structured data over a range of scales. The effectiveness of these new approaches is demonstrated by some novel results obtained in the analysis of brain artery trees. The scale space analysis reveals a deeper relationship between structure and age. These methods are the first to find a statistically significant gender difference.

email: shen.unc@gmail.com

7j. MULTICATEGORY LARGE-MARGIN UNIFIED MACHINES

Chong Zhang*, University of North Carolina, Chapel Hill
Derek Y. Chiang, University of North Carolina, Chapel Hill
Yufeng Liu, University of North Carolina, Chapel Hill

Hard and soft classifiers are two important groups of techniques for classification problems. Logistic regression and Support Vector Machines are typical examples of soft and hard classifiers respectively. The essential difference between these two groups is whether one needs to estimate the class conditional probability for the classification task or not. In practice it is unclear which one to

use in a given situation. To tackle this problem, the Large-margin Unified Machine (LUM) was recently proposed as a unified family to embrace both groups. The LUM family enables one to study the behavior change from soft to hard binary classifiers. For multicategory cases, however, the concept of soft and hard classification becomes less clear. In that case, class probability estimation becomes more involved as it requires estimation of a probability vector. In this paper, we propose a new Multicategory LUM (MLUM) framework to investigate the behavior of soft versus hard classification under multicategory settings. Our theoretical and numerical results help to shed some light on the nature of multicategory classification and its transition behavior from soft to hard classifiers. The numerical results suggest that the proposed MLUM is highly competitive among several existing methods.

email: chongz@live.unc.edu

8. POSTERS: MODEL, PREDICTION, VARIABLE SELECTION AND DIAGNOSTIC TESTING

8a. PENALIZED COX MODEL FOR IDENTIFICATION OF VARIABLES' HETEROGENEITY STRUCTURE IN POOLED STUDIES

Xin Cheng*, New York University School of Medicine
Wenbin Lu, North Carolina State University
Mengling Liu, New York University School of Medicine

Pooled analysis that pools datasets from multiple studies and treats them as one large single set of data can achieve large sample size to allow increased power to investigate variables' effects, if homogeneous across studies. However, inter-study heterogeneity often exists in pooled studies, due to differences such as study population and sample collection method. To evaluate variables' homogeneous and heterogeneous structure and estimate their effects, we propose a penalized partial likelihood approach with an adaptively weighted L1 penalty on variable's average effects and a combination of adaptively weighted L1 and L2 penalty on heterogeneous effects. We show that our method can identify the structure of variables as heterogeneous effects, nonzero homogeneous effects and null effects, and give consistent estimation of variables' effects simultaneously. Furthermore, we extend our method to high dimension situation, where the number of parameters diverges with sample size. The proposed selection and estimation procedure can be easily implemented using the iterative shooting algorithm. We conduct extensive numerical studies to evaluate the practical performance of the proposed method and demonstrate it using two real studies.

email: zhichi1@gmail.com

8b. BAYESIAN PREDICTIVE DIVERGENCE BASED MODEL SELECTION CRITERIA FOR CENSORED AND MISSING DATA

Liwei Wang*, North Carolina State University
Sujit K. Ghosh, North Carolina State University

Bayesian model selection for data analysis becomes a challenging task when observations are subject to data irregularities like censoring and missing values. Often a linear mixed effects framework is used to approximate semiparametric models, but some of the popular Bayesian criteria like DIC do not work well in choosing among mixed models and there have been ambiguities in defining the deviances for mixed effect models. To illustrate the proposed model selection criteria, first we develop a flexible class of mixed effects models based on a sequence of Bernstein polynomials with varying degrees and propose a predictive divergence based model selection criterion for the fully observed data. We then extend the model selection criteria to accommodate the data irregularities and develop an importance sampling based MCMC method to compute the criteria. Various simulated data scenarios are used to compare the performance of the proposed model selection methodology with some of the popular Bayesian model selection methodologies. The newly proposed models and associated model selection criteria are also illustrated using real data analysis.

email: lwang16@ncsu.edu

8c. A TUTORIAL ON LEAST ANGEL REGRESSION

Wei Xiao*, North Carolina State University
Yichao Wu, North Carolina State University
Hua Zhou, North Carolina State University

The least angel regression (LAR) was proposed by Efron, Hastie, Johnstone and Tibshirani (2004) for continuous model selection in linear regression. It is motivated by a geometric argument and tracks a path along which the predictors enter successively and the active predictors always maintain the same correlation (angle) with the residual vector. Although gaining popularity quickly, extensions of LAR seem rare compared to the penalty methods. In this expository article, we show that the powerful geometric idea of LAR can be extended in a fruitful way. We propose a ConvexLAR algorithm that works for any convex loss function and naturally extends to group selection and data adaptive variable selection. Variable selection in recurrent event and panel count data analysis, Ada-Boost, and Gaussian graphical model, is reconsidered from the ConvexLAR angle.

email: wxiao@ncsu.edu

8d. VARIABLE SELECTION IN MEASUREMENT ERROR MODELS VIA LEAST SQUARES APPROXIMATION

Guangning Xu*, North Carolina State University
Leonard A. Stefanski, North Carolina State University

A fundamental problem in biomedical research is identifying key risk factors and determining their impact on health outcomes via statistical modeling. Due to device limitations and within-subject variation, some risk factors are measured with error, e.g., blood pressure. Ignoring measurement error adversely impacts variable selection and model fitting and thus complicates the statistical modeling. When measurement error is present, popular variable selection methods, such as LASSO, ALASSO and SCAD are not appropriate. We propose a new method for variable selection in measurement error models by integrating well-established measurement error modeling methods with the least squares approximation (LSA) variable selection method of Wang and Leng (2007). The resulting estimators are consistent and asymptotically normal in the usual case that the measurement error corrected estimator is root-n consistent. The method inherits the oracle property when an adaptive penalty is used and the tuning parameter is well selected. The key advantage of our new method is that it provides a unified solution to the variable selection in measurement error models and greatly eases computing by using existing algorithms.

email: gxu@ncsu.edu

8e. SMOOTHED STABILITY SELECTION FOR ANALYSIS OF SEQUENCING DATA

Eugene Urrutia*, University of North Carolina, Chapel Hill
Yun Li, University of North Carolina, Chapel Hill
Michael C. Wu, University of North Carolina, Chapel Hill

High dimensional data are increasingly common. Difficulties in model interpretation and limited power to detect effects have led to the use of variable selection methods, including penalized regression methods such as the LASSO. Recent advances in the variable selection literature suggest that resampling strategies which include stability selection, complementary stability selection, and bolasso, can offer improvements over the LASSO and non-resampling based approaches. We show that common resampling based methods can be recast as LASSO with reweighted observations where weights are discrete and identically distributed from a specified distribution. Sequencing data presents an additional challenge in that many predictors are rare (minor alleles observed in only a few individuals) and have a high probability of exclusion in the resampling schemes. Thus, we have developed the smooth stability selection procedure where we replace

the discrete weights with continuous weights, and thus avoid excluding rare predictors. Simulation results and real data analyses suggest that our proposed method increases power to select rare variables while retaining type I error control.

email: gene.urrutia@gmail.com

8f. JOINT MODELING OF TIME-TO-EVENT DATA AND MULTIPLE RATINGS OF A DISCRETE DIAGNOSTIC TEST WITHOUT GOLD STANDARD

Seung Hyun Won*, University of Pittsburgh
Gong Tang, University of Pittsburgh
Ruosha Li, University of Pittsburgh

Histologic tumor grade is a strong predictor of risk of recurrence in breast cancer. However, tumor grade readings by pathologists are susceptible to intra- and inter- observer variability due to its subjective nature. For this limitation, tumor grade is not included in the breast cancer staging system. Latent class models are considered for analysis of such discrete diagnostic tests with the underlying truth as a latent variable. However, the model parameters are only locally identifiable that any permutation on the categories of the truth also leads to the same likelihood function. In many circumstances, the underlying truth is known associated with risk of certain event in a trend. Here we propose a joint model with a Cox proportional hazard model for the time-to-event data where the underlying truth is a latent predictor. The joint model not only fully identifies all model parameters but also provide valid assessment of the association between the diagnostic test and the risk of event. The EM algorithm was used for estimation. We showed that the M-steps are equivalent to fitting survey-weighted Cox models. The proposed method is illustrated in the analysis of data from a breast cancer clinical trial and simulation studies.

email: sew53@pitt.edu

8g. LOGIC REGRESSION MODELING WITH REPEATED MEASUREMENT DATA AND ITS APPLICATIONS ON SYNDROMIC DIAGNOSIS OF VAGINAL INFECTIONS IN INDIA

Tan Li*, Florida International University
Wensong Wu, Florida International University

Most of regression methodologies are unable to find the effect of complex interaction but only simple interactions (two-way or three-way). However, the complex interaction between more than three predictors may be the cause to the differences in response, especially when all the predictors are binary. Logic regression, developed by Ruczinski and LeBlanc (2003), is a generalized regression methodology, which has been used to construct the complex interactions between binary predictors as Boolean logic statements. However, this methodology is not applicable to the repeated measurement data, which is the common data type in the public health field. The purpose of this paper is going to study the logic regression modeling with repeated measurement data. The proposed method will be compared with the logic regression without considering repeated measurement on simulated data. The proposed method will be also applied to the real data for the syndromic diagnosis of vaginal infections in India and compared to a commonly used algorithm developed by the World health Organization (WHO).

email: tanli@fiu.edu

8h. SEQUENTIAL CHANGE POINT DETECTION IN LINEAR QUANTILE REGRESSION MODELS

Mi Zhou*, North Carolina State University
Huixia (Judy) Wang, North Carolina State University

Sequential detection of change point has many important applications in finance, econometrics, engineering etc, where it is desirable to raise an alarm as soon as the system or model structure has an abrupt change. In the current literature, most methods for sequential monitoring focus on the mean function. We develop a new sequential change point detection method for linear quantile regression models. The proposed statistic is based on the cusum of quantile score functions. The method can be used to detect change points at a single quantile level or across quantiles, and can accommodate both homoscedastic and heteroscedastic errors. We establish the asymptotic properties of the developed test procedure, and assess its finite sample performance by comparing it to existing change point detection methods for linear regression models.

email: mzhou2@ncsu.edu

8i. ASSESSMENT OF THE CLINICAL UTILITY OF A PREDICTIVE MODEL FOR COLORECTAL ADENOMA RECURRENCE

Mallorie Fiero*, University of Arizona
Dean Billheimer, University of Arizona
Joshua Mallet, University of Arizona
Bonnie LaFleur, University of Arizona

Current methods for evaluating predictive models include evaluating sensitivity, specificity and area under the ROC curve (AUC). Other metrics, such as the Brier score, Somers' Dxy, and R^2 are also advocated by statisticians. The potential limitation of all of these predictive assessments is the translation from statistical to clinical relevance. Clinical relevance includes individual determination of risks of treatments, as well as adoption of novel markers (biomarkers) to determine prognostic outcome. In this paper, we evaluate two potential clinical metrics of prediction performance, the predictiveness curve, proposed by Pepe, et. al. (2008), and decision curve analysis, proposed by Vickers, et. al. (2006). We apply these methodologies to a sample of patients diagnosed with colorectal adenomatous polyps, a precursor lesion for colorectal cancer, and evaluate their risk of polyp recurrence. Baseline pathologic, patient, and molecular characteristics are included in the predictive model, and we examine the benefits and drawbacks of each clinical decision support tool.

email: mfiero@email.arizona.edu

8j. ASSESSING ACCURACY OF POPULATION SCREENING USING LONGITUDINAL MARKER

Paramita Saha-Chaudhuri*, Duke University
Patrick Heagerty, University of Washington

The World Health Organization defines screening as the presumptive identification of unrecognized disease or defects by means of tests, examinations or other procedures that can be applied rapidly. With the progress of medicine in recent decades, there is now considerable focus on early detection of disease via population screening, giving the patient more treatment options if the disease is found by screening and consequently a better prognosis outlook. Both the benefits and harms of screening are well-documented. There is considerable debate as to whether early screening for diseases, such as cancer and cardiovascular disease, is useful, has a positive trade-off and has a broad public health impact. Once a screening protocol is introduced in practice, there is much controversy as to whether it is possible to forgo screening at all. Given that a screening test is introduced, it is imperative to assess the accuracy of the screening protocol. We introduce a new approach, Screening ROC, that can be used to assess the accuracy of a screening protocol. We demonstrate the approach with simulated and real dataset.

email: paramita.sahachaudhuri@duke.edu

8k. ASSESSING CALIBRATION OF RISK PREDICTION MODELS FOR POLYTOMOUS OUTCOMES

Kirsten Van Hoorde*, Katholieke Universiteit, Leuven, Belgium
Sabine Van Huffel, Katholieke Universiteit, Leuven, Belgium
Dirk Timmerman, Katholieke Universiteit, Leuven, Belgium
Ben Van Calster, Katholieke Universiteit, Leuven, Belgium

Risk prediction models assist clinicians in making treatment decisions. Therefore the estimated risks should correspond to observed risks (calibration). For binary outcomes, tools to assess calibration exist, e.g. calibration-in-the-large, calibration slope, and calibration plots. We extend these tools to models for nominal outcomes developed using baseline-category logistic regression. The logistic recalibration model is a baseline-

category fit of outcome Y with k categories, in which each category i (i=2,..., k) is compared to reference category 1 using log[P(Y=i)/P(Y=1)] = a_i+b_i*lp_i, with lp_i the linear predictor for category i versus 1. Thus, each equation contains only the linear predictor of the category involved. Calibration-in-the-large (a_i|b_i=1) assesses whether risks are too high or low, and calibration slopes (b_i) whether risks are over- or underfitted. A parametric calibration plot uses the logistic recalibration model, thus assuming linearity. A non-parametric alternative estimates a'_i+b'_i*s(lp_i), with s() a vector spline, using a vector generalized additive model. A case study is presented on the diagnosis of ovarian tumors as benign, borderline or invasive.

email: kirsten.vanhoorde@esat.kuleuven.be

8l. TESTING MULTIPLE BIOLOGICAL MEDIATORS SIMULTANEOUSLY

Simina M. Boca*, National Cancer Institute, National Institutes of Health
Rashmi Sinha, National Cancer Institute, National Institutes of Health
Amanda J. Cross, National Cancer Institute, National Institutes of Health
Steven C. Moore, National Cancer Institute, National Institutes of Health
Joshua N. Sampson, National Cancer Institute, National Institutes of Health

Numerous statistical methods adjust for "multiple comparisons" when testing whether multiple biomarkers, such as hundreds or thousands of gene expression or metabolite levels, are directly associated with an outcome. However, other than the simple Bonferroni correction, there are no adjustment methods for testing whether multiple biomarkers are biological mediators between a known risk factor and a disease. We propose a novel permutation test which controls the Family Wise Error Rate (FWER) when testing multiple mediators. The composite null hypothesis must allow for each potential mediator to be linked with either exposure or outcome, just not both. This requires our novel two-step permutation procedure which considers each association separately. We evaluated the statistical power of our method via simulation. We also applied it to a case/control study examining whether any of 167 serum metabolites mediate the relationship between dietary risk factors, specifically red-meat and fish consumption, and colorectal adenoma, a precursor of cancer.

email: york60@gmail.com

8m. MARGINAL ANALYSIS OF MEASUREMENT AGREEMENT AMONG MULTIPLE RATERS WITH NON-IGNORABLE MISSING RATINGS

Yunlong Xie*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Zhen Chen, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

In diagnostic medicine, several measurements have been developed to evaluate the agreements among raters when the data are complete. In practice, raters may not be able to give definitive ratings to some participants because symptoms may not be clear cut. Simply removing subjects with missing ratings may produce biased estimates and result in loss of efficiency. In this article, we propose a within-cluster resampling (WCR) procedure and a marginal approach to handle non-ignorable missing data in measurement agreement data. Simulation studies show that both WCR and marginal approach provide unbiased estimates and have coverage probabilities close to the nominal level. The proposed methods are applied to the data set from the Physician Reliability Study in diagnosing endometriosis.

email: yunlong.xie@nih.gov

8n. IMPROVING CLASSIFICATION ACCURACY BY COMBINING LONGITUDINAL MEASUREMENTS WHEN ANALYZING CENSORED BIOMARKER DATA DUE TO A LIMIT OF DETECTION

Yeonhee Kim*, Gilead Sciences
Lan Kong, Penn State Hershey College of Medicine

Diagnostic or prognostic performance of a biomarker is commonly assessed by the area under the receiver operating characteristic curve. It is expected that longitudinal information rather than using single time point measurement would lead to better performance. However, incorporating repeated measurements in the ROC analysis is not straightforward, and the evaluation may be even complicated by the limitation of technical accuracy that causes biomarker measurements being left or right censored at detection limits. Ignorance of correlated nature of longitudinal data and censoring issue may yield spurious estimation of AUC, and could rule out potentially informative biomarkers from further investigation. We introduce a linear combination of longitudinal measurements under the goal of optimizing AUC, while accounting for censored observations. Investigators are able to not only evaluate the potential classification power of a marker, but also determine relative importance of each time point in the clinical decision making process. Our method is assessed in the simulation study and illustrated using a real data set collected from hospitalized patients with community acquired pneumonia.

email: yhkimbani@gmail.com

8o. A HYBRID BAYESIAN HIERARCHICAL MODEL COMBINING COHORT AND CASE-CONTROL STUDIES FOR META-ANALYSIS OF DIAGNOSTIC TESTS: ACCOUNTING FOR DISEASE PREVALENCE AND PARTIAL VERIFICATION BIAS

Xiaoye Ma*, University of Minnesota
Haitao Chu, University of Minnesota
Yong Chen, University of Texas
Stephen R. Cole, University of North Carolina, Chapel Hill

Bivariate random effects models have been recommended to jointly model sensitivities and specificities in meta-analysis of diagnostic tests accounting for between-study heterogeneity. Because the severity and definition of disease may differ among studies due to the design and the population, the sensitivities and specificities of a diagnostic test may depend on disease prevalence. To account for the potential dependence, trivariate random effects models have been proposed. However, this approach can only include cohort studies with information for study-specific disease prevalence. In addition, some diagnostic accuracy studies select a subset of samples based on test results to be verified. It is known that ignoring unverified subjects can lead to partial verification bias in the estimation of accuracy indices in single study. However, the impact of this bias on meta-analysis of diagnostic tests has not been investigated. As many diagnostic accuracy studies use case-control designs, we propose a hybrid Bayesian hierarchical model combining cohort and case-control studies to account for prevalence and to correct partial verification bias at the same time. We investigate the performance of the proposed methods through a set of simulation studies, and present a case study on assessing the diagnostic accuracy of MRI in detecting lymph node metastases.

email: maxxx372@umn.edu

9. POSTERS: ENVIRONMENTAL, EPIDEMIOLOGICAL AND HEALTH SERVICES STUDIES

9a. ESTIMATING SOURCE-SPECIFIC EFFECTS OF FINE PARTICULATE MATTER EMISSIONS ON CARDIOVASCULAR AND RESPIRATORY HOSPITALIZATIONS USING SPECIATE AND NEI DATA

Amber J. Hackstadt*, Johns Hopkins University
Roger D. Peng, Johns Hopkins University

Previous work has established an association between exposure to fine particulate matter (PM) and the risk for mortality and morbidity with evidence that the health effects vary for the different chemical constituents that compose fine PM. This suggests that sources of fine PM have varying effects on health due to their differences in the chemical constituents that contribute to their emis-

sions and their differences in the contributions of each chemical constituent. Thus, obtaining source-specific health effects estimates can help better regulate the threat to public health. We propose a Bayesian source apportionment model to apportion ambient fine PM measurements into source-specific contributions that incorporates information about source emissions from two United States Environmental Protection Agency databases, SPECIATE and the National Emissions Inventory (NEI). These source contributions are incorporated into time series regression models to obtain estimates of short-term risks of these sources on cardiorespiratory hospital admissions. The proposed models are used to obtain source-specific health effect estimates for cardiovascular and respiratory hospital admissions for Boston, Massachusetts and Phoenix, Arizona.

email: ahacksta@jhsph.edu

9b. THE IMPACT OF VALUES BELOW THE MINIMUM DETECTION LIMIT ON SOURCE APPORTIONMENT RESULTS

Jenna R. Krall*, Johns Hopkins University
Roger D. Peng, Johns Hopkins University

Studies of particulate matter sources frequently apply factor analysis methods to chemical constituent data to identify sources and obtain source concentrations, referred to as source apportionment. Constituents with low mean concentrations often have daily values that fall below the minimum detection limit (MDL) and these missing values may impact source apportionment results. We examined how values below the MDL affect source apportionment by comparing results between simulated chemical constituent data with and without missing values below the MDL. We imputed values below the MDL using three methods: (1) applying 1/2*MDL, (2) eliminating chemical constituents with a large percentage of values below the MDL, and (3) applying a novel method using draws from a multivariate truncated lognormal distribution. Across all imputation methods, identifying source types and recovering source concentration distributions were more difficult as the number of true underlying sources increased and as the number of chemical constituents with values below the MDL increased. Dropping constituents with a large percent of values below the MDL performed worse than applying a multivariate truncated lognormal distribution or 1/2*MDL. When chemical constituent data have concentrations below the MDL, source apportionment results may be heavily impacted.

email: jkrall@jhsph.edu

9c. MODELING THE DYNAMIC RELATIONSHIPS AMONG AIR POLLUTANTS OVER TIME AND SPACE THROUGH PENALIZED SPLINES

Zhenzhen Zhang*, University of Michigan
Brisa N. Sanchez, University of Michigan

The latent factor model is a popular approach to modeling relationships among multiple pollutants. But the time-invariant factor structure cannot detect the change of relationship over time. In this paper we develop a factor model that uses penalized splines to model the time-variant factor loadings. In addition, factor loadings are also allowed to vary across space to capture and compare the pollution level between different regions.

email: zhzh@umich.edu

9d. BAYESIAN COMPARATIVE CALIBRATION OF SELF-REPORTS AND PEER-REPORTS OF ALCOHOL USE ON A SOCIAL NETWORK

Miles Q. Ott*, Brown University
Joseph W. Hogan, Brown University
Krista J. Gile, University of Massachusetts
Crystal D. Linkletter, Brown University
Nancy P. Barnett, Brown University

Analysis of social network data is central to understanding how health related behaviors are influenced by the social environment. Methods for network analysis are particularly relevant in substance use research because substance use is influenced by the behaviors and attitudes of peers. The UrWeb study collected data on alcohol in a social network formed by college students living in a freshman dormitory. By using two imperfect sources of information collected on the network (self-reported alcohol consumption and peer-reported alcohol consumption), rather than solely self-reports or peer-reports of alcohol consumption, we are able to gain insight into alcohol consumption on both the population and the individual level, as well as information on the bias of individual peer-reports. In order to carry out this analysis, we develop a novel Bayesian comparative calibration model that characterizes the joint distribution of both self and peer-reports on the network for estimating peer-reporting bias in network surveys, and apply this to the UrWeb data.

email: miles_ott@brown.edu

9e. DIFFERENTIAL IMPACT OF JUNK FOOD POLICIES ON POPULATION CHILDHOOD OVERWEIGHT TRENDS BY SOCIO-ECONOMIC STATUS

Sarah Abraham*, University of Michigan
Brisa N. Sanchez, University of Michigan
Emma V. Sanchez-Vaznaugh, San Francisco State University
Jonggyu Baek, University of Michigan

Policies restricting sales of competitive ("junk") foods and beverages in schools have been previously shown to influence population-level time-trends in the percent of overweight 5th and 7th grade children in the state of California. The percent of overweight children increased from year to year prior to enactment of policies restricting sales of junk food, but after the policies were implemented, the percentage overweight either flattened or decreased over time depending on sex and grade of the children. The extent of these results across school-level socio-economic status (SES) remains unclear. It is hypothesized that children attending schools with higher economic resources would have the highest benefit compared to children in lower SES schools. Thus, data was stratified further by school-level SES, determined based on the percentage of residents near the school with attained bachelors' degrees. We used generalized linear mixed models to test if the rate of change in the proportion of overweight children significantly differed according to school-level SES status, while accounting for clustering of children within schools. We found that both males and females in 5th and 7th grade who were in the highest SES stratum experienced a decline in the population-level trend of overweight children; however, results were mixed for the medium and low SES strata.

email: sabraha@umich.edu

9f. ASSOCIATION OF SELECTED HEAVY METALS AND FATTY ACIDS WITH OBESITY

Stephanie Schilz*, Concordia College
Budhinath Padhy, South Dakota State University
Douglas Armstrong, South Dakota State University
Gemechis Djira, South Dakota State University

Obesity, a serious and costly health issue, has become more prevalent in the United States in the recent decades. Studies have linked different heavy metals to weight, obesity, and body mass index (BMI). Selected fatty acids have also been found to have associations with BMI. To study this further, National Health and Nutrition Examination Survey (NHANES) data was used. NHANES, conducted by the National Center for Health Statistics, uses a complex survey design, yielding a representative sample of the US non-institutionalized civilian population. Six heavy metals and twenty four different fatty acids were examined, with age, gender, and race used as covariates, as well as

creatinine to account for kidney damage caused by the heavy metals. A survey logistic regression was used for the analysis with the binary response being obesity. A statistical issue in this analysis is the use of weights coming from different subsamples in the survey data. Barium was found to be the only significant heavy metal in the final model, and myristoleic acid, eicosadienoic acid, stearic acid, gamma-linolenic acid, alpha-linolenic acid, and docosahexaenoic acid were the significant fatty acids. The significance of myristoleic acid, eicosadienoic acid, and stearic acid were consistent with existing literature.

email: gemechis.djira@sdstate.edu

9g. A SEMI-NONPARAMETRIC PROPENSITY SCORE MODEL FOR TREATMENT ASSIGNMENT HETEROGENEITY WITH APPLICATION TO ELECTRONIC MEDICAL RECORD DATA

Baiming Zou*, University of North Carolina, Chapel Hill
Fei Zou, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Haibo Zhou, University of North Carolina, Chapel Hill

Analyzing electronic medical record (EMR) data to compare the effectiveness of different treatments is a central component in the moment of comparative effectiveness research (CER). A key statistical challenge in this research is how to properly analyze the EMR data, which is different from the clinical trial data where the treatment assignment is random, to unbiasedly and efficiently estimate the true treatment effect in the real world large-scale medical data. Existing methods, such as propensity score approach, generally assume all confounding variables are observed. As clinical dataset for CER research are not designed to capture all confounding variables, heterogeneity will exist in the real world EMR or clinical dataset. Most importantly, it is known that real world patient's treatment assignment is influenced by the physician (care provider), the system (e.g. insurance type), and the patient themselves (e.g. religion). Hence, heterogeneity in the treatment assignment in real world EMR or clinical data needs to be taken into account in estimating the true treatment effect. We propose a semi-nonparametric propensity score (SNP-PS) model to deal with the heterogeneity of treatment assignment. Our model makes no specific distribution assumption on the random effects except that the distribution function is smooth, and thus is more robust to model misspecifications. A truncated Hermite polynomial along with the normal density is used to approximate the unknown density of the heterogeneity. To avoid the potential large Monte Carlo errors of sampling based algorithms, we developed an adaptive EM algorithm for SNP-PS parameter estimates. More importantly, a robust and consistent variance estimate for the parameter estimator is proposed which corrects the bias of naive variance estimation from the second stage regression model that is routinely used in practice. This finding is critical in practice, as it will greatly help to reduce both false positive and negative rates of those studies where naive variance estimation is used. We established the asymptotic results for the treatment effect estimator.

email: bzou@email.unc.edu

9h. ESTIMATING INCREMENTAL COST-EFFECTIVENESS RATIOS AND THEIR CONFIDENCE INTERVALS WITH DIFFERENT TERMINATING EVENTS FOR SURVIVAL TIME AND COSTS

Shuai Chen*, Texas A&M University
Hongwei Zhao, Texas A&M University

Cost-effectiveness analysis is an important component of the economic evaluation of new treatment options. In many clinical and observational studies of costs, censored data pose challenges to the cost-effectiveness analysis. We consider a special situation where the terminating events for survival time and costs are different. Traditional methods for statistical inference offer no means for dealing with censored data in these circumstances. To address this gap, we propose a new method for deriving the confidence interval for this incremental cost-effectiveness ratio, based on the counting process and the general theory for missing data process. The simulation studies and real data example show that our method performs very well for some practical settings, revealing a great potential for application to actual settings in which terminating events for survival time and costs differ.

email: shuai@stat.tamu.edu

9i. A GENERAL FRAMEWORK FOR SENSITIVITY ANALYSIS OF COST DATA WITH UNMEASURED CONFOUNDING

Elizabeth A. Handorf*, Fox Chase Cancer Center
Justin E. Bekelman, University of Pennsylvania
Daniel F. Heitjan, University of Pennsylvania
Nandita Mitra, University of Pennsylvania

Estimates of treatment effects on cost from observational studies are subject to bias if there are unmeasured confounders. It is therefore advisable in practice to assess the potential magnitude of such biases; in some cases, closed-form expressions are available. We derive a general adjustment formula using the moment-generating function for log-linear models and explore special cases under plausible assumptions about the distribution of the unmeasured confounder. We assess the performance of the adjustment by simulation, in particular examining robustness to a key assumption of conditional independence between the unmeasured and measured covariates given the treatment indicator. We show how our method is applicable to cost data with informative censoring, and apply our method to SEER-Medicare cost data for a stage II/III muscle-invasive bladder cancer cohort. We evaluate the costs for radical cystectomy vs. combined radiation/chemotherapy, and find that the significance of the treatment effect is sensitive to unmeasured Bernoulli, Poisson, and Gamma confounders.

email: elizabeth.handorf@fccc.edu

9j. THE APPROPRIATENESS OF COMORBIDITY SCORES TO ACCOUNT FOR CLINICAL PROGNOSIS AND CONFOUNDING IN OBSERVATIONAL STUDIES

Brian L. Egleston*, Fox Chase Cancer Center
Steven R. Austin, Johns Hopkins University
Yu-Ning Wong, Fox Chase Cancer Center
Robert G. Uzzo, Fox Chase Cancer Center
J. R. Beck, Fox Chase Cancer Center

Comorbidity adjustment is an important goal of health services research and clinical prognosis. When adjusting for comorbidities in statistical models, researchers can include comorbidities individually or through the use of summary measures such as the Charlson Comorbidity Index or Elixhauser score. While many health services researchers have compared the utility of comorbidity scores using data examples, there has been a lack of mathematical rigor in most of the evaluations. In the statistics literature, Hansen (Biometrika 2008) provided a theoretical justification for the use of prognostic scores. We examined the conditions under which individual versus summary measures are most appropriate. We expand on Hansen's work, and show that comorbidity scores created analogously to the Charlson Comorbidity Index are indeed appropriate balancing scores for prognostic modeling and comorbidity adjustment.

email: Brian.Egleston@fccc.edu

9k. ASSESSMENT OF HEALTH CARE QUALITY WITH MULTILEVEL MODELS

Christopher Friese*, University of Michigan
Rong Xia, University of Michigan
Mousumi Banerjee, University of Michigan

Assessment of health care quality provided in US hospitals is not only an important medical research target but also a challenging statistics problem. To account for the hierarchical structure in the data, we applied multilevel logistic models to study the effects of hospital characters on health care quality, specially the risk-adjusted mortality and failure-to-rescue. We have found that patients treated in the Magnet hospitals have significant lower rate of mortality and failure-to-rescue. We have also compared the Magnet hospitals to non-Magnet hospitals to discover the differences in hospital size, nursing, cost, location etc. These conclusions were based on the national wide data from year 1998 to 2008.

email: rongxia@umich.edu

9l. TESTING FOR BIASED INFERENCE IN CASE-CONTROL STUDIES

David Swanson*, Harvard University
Rebecca A. Betensky, Harvard University

Survival bias is a long-recognized problem in case-control studies, and many varieties of bias can come under this umbrella term. We focus on one of them, termed Neyman's bias or "prevalence-incidence bias." It occurs in case-control studies when exposure affects both disease and disease-induced mortality, and we give a formula for the observed, biased odds ratio under such conditions. We compare our result with previous investigations into this phenomenon and consider models under which this bias may or may not be significant. Finally, we propose three hypothesis tests to identify when Neyman's bias may be present in case-control studies. We apply these tests to two data sets, one of stroke mortality and another of brain cancer, and find some evidence of Neyman's bias in both cases.

email: dswanson@fas.harvard.edu

10. POSTERS: NON-PARAMETRIC AND SPATIAL MODELS

10a. NONPARAMETRIC MANOVA APPROACHES FOR MULTIVARIATE OUTCOMES IN SMALL CLINICAL STUDIES

Fanyin He*, University of Pittsburgh
Sati Mazumdar, University of Pittsburgh

Comparisons between treatment groups play a central role in clinical research. As these comparisons often entail correlated dependent variables, the multivariate general linear model has been accepted as a standard tool. However, parametric multivariate methods require assumptions such as multivariate normality. Standard statistical packages delete subjects when data are not available for some of the dependent variables. This article addresses the issues of missing data and violation of multivariate normality assumption. It focuses on the multivariate Kruskal-Wallis (MKW) test for group comparisons. We have written an R-based program that can do the MKW test, likelihood-based or permutation-based. Simulation studies are done under different scenarios to compare the performances of MKW test and classical MANOVA tests. We consider skewed continuous and ordinal correlated outcomes, and outcomes with different patterns and amounts of missingness. Simulation studies show that the nonparametric methods have reasonable control of type I error. Statistical issues related to sample size calculations for the detection of effect sizes in a multivariate set up are discussed.

email: fah11@pitt.edu

10b. NONPARAMETRIC REGRESSION FOR EVENT TIMES IN MULTISTATE MODELS WITH CLUSTERED CURRENT STATUS DATA WITH INFORMATIVE CLUSTER SIZE

Ling Lan*, Georgia Health Sciences University
Dipankar Bandyopadhyay, University of Minnesota
Somnath Datta, University of Louisville

We consider data sets containing current status information on individual units each undergoing a multistate system. The state processes have a clustering structure such that the size of each cluster is random and may be in correlation with common characteristics of the multistate models in the cluster. We propose nonparametric estimators for the state occupation probabilities at a given time conditional on a continuous and a discrete covariate. Weighted monotonic regression and smoothing are used to uniquely define the state occupation probability regression estimators. A detailed simulation study evaluated the global performance of the proposed nonparametric estimator. An illustrative application to a dental disease data is also presented.

email: llan@georgiahealth.edu

10c. POPULATION SPATIAL CLUSTERING TO DETERMINE OPTIMAL PLACEMENT OF CARE CENTERS

John R. Zeiner*, University of Pittsburgh Medical Center
Jennie Wheeler, University of Pittsburgh Medical Center

Background: A regional health plan identified a potential cost savings by internal care optimization logic to identify claims that could be optimized to an urgent care setting or a retail care setting. Geodetic distance between the existing network and the health plan members were calculated. The potential savings was halved because of the membership living outside the ideal radius. Locations needed to be identified for network expansion. Method: Geodetic distance between street-level geocoded members' residence to the care centers were calculated. Those members outside of 5 miles (retail care) or 10 miles (urgent care) radius to the closest center were excluded in the analysis. Longitude / latitude were converted to the Euclidean space for cluster analysis. Disjoined clusters were calculated using k-means modeling. The seeds were the center of the population density. To minimize the effect of outliers, each county was modeled separately. An adaptive algorithm was written to minimize the number of clusters needed to cover the population. The seeds coordinates were converted back to geographic coordinate system. Geodetic distance was calculated to confirm optimization. Seed coordinates were evaluated in Google Maps©. Results: Ideal locations for expansion or partnerships were identified to provide increased access to care options.

email: zeinerj@upmc.edu

10d. PROCESS-BASED BAYESIAN MELDING OF OCCUPATIONAL EXPOSURE MODELS AND INDUSTRIAL WORKPLACE DATA

Joao Monteiro*, Duke University
Sudipto Banerjee, University of Minnesota
Gurumurthy Ramachandran, University of Minnesota

In industrial hygiene a worker's exposure to chemical, physical and biological agents is increasingly being modeled using deterministic physical models. However, predicting exposure in real workplace settings is challenging and approaches that simply regress on a physical model (e.g. straightforward non-linear regression) are less effective as they do not account for biases attributable, at least in part, to extraneous variability. This also impairs predictive performance. We recognize these limitations and provide a rich and flexible Bayesian hierarchical framework, which we call process-based Bayesian melding (PBBM), to synthesize the physical model with the field data. We reckon that the physical model, by itself, is inadequate for enhanced inferential performance and deploy (multivariate) Gaussian processes to capture extraneous uncertainties and underlying associations. We propose rich covariance structures for multiple outcomes using latent stochastic processes.

email: joao.monteiro@stat.duke.edu

10e. A GENERALIZED WEIGHTED REGRESSION APPROACH FOR ASSESSING LOCAL DETERMINANTS IN THE PREDICTION OF RELIABLE COST ESTIMATES

Giovanni Migliaccio, University of Washington
Michele Guindani, University of Texas MD Anderson Cancer Center
Maria Incognito, Department of Civil, Environmental, Building and Chemical Engineering, Bari, Italy
Linlin Zhang*, Rice University

The availability of reliable cost estimates is fundamental for the successful implementation of a variety of health related programs, including construction of new facilities. In addition to correctly predicting global trends that may impact the cost estimates, e.g. general cost increases for some of the materials, a number of factors contribute to variability of the cost estimates according to local or regional characteristics. In practice, a very common approach for performing quick-order-of-magnitude estimates is based on using location adjustment factors (LAFs) that compute historically based costs by project location. This research provides a contribution to the body of knowledge by investigating the relationships between

a commonly used set of LAFs, the City Cost Indexes (CCI) by RSMeans, and the socio-economic variables included in the ESRI Community Sourcebook. We use a Geographically Weighted regression analysis (GWR) and compare the results with interpolation methods, which are more common in the industry. We show that GWR is the most appropriate way to model the local variability of cost estimates on the proposed dataset, and we assess how the effect of each single covariate varies from state to state.

email: lz17@rice.edu

10f. EVALUATION OF NON-PARAMETRIC PAIR CORRELATION FUNCTION ESTIMATE FOR LOG-GAUSSIAN COX PROCESSES UNDER INFILL ASYMPTOTICS

Ming Wang*, Emory University
Jian Kang, Emory University
Lance A. Waller, Emory University

Log-Gaussian Cox Processes (LGCPs), Cox point processes where the log intensity function follows a Gaussian Process, provide a very flexible framework for modeling heterogeneous spatial point processes. The pair correlation function (PCF) plays a vital role in characterizing second-order spatial dependencies in LGCPs and delivers key input on association structures, yet empirical estimation of the PCF remains challenging, even more so for spatial point processes in one dimension (points along a line). We consider two common approaches for edge-correction of nonparametric PCF estimates in two dimensions and evaluate their performance through theory and simulation when applied to one-dimensional data. Our results reveal that an algorithm based on theoretical formulae combined with finite sample simulation of the k^th (k<=4) moment provides superior performance over a standard non-parametric approach. In addition, we propose a new edge-correction method for one-dimensional point processes providing improvement over current methods with respect to bias reduction.

email: wm_pku@hotmail.com

11. SPATIAL STATISTICS FOR ENVIRONMENTAL HEALTH STUDIES

A BAYESIAN SPATIALLY-VARYING COEFFICIENTS MODEL FOR ESTIMATING MORTALITY RISKS ASSOCIATED WITH THE CHEMICAL COMPOSITION OF FINE PARTICULATE MATTER

Francesca Dominici*, Harvard School of Public Health

Although there is a large epidemiological literature describing the chronic effects associated with long-term exposure to particulate matter (PM2.5), evidence regarding the toxicity of the individual chemical constituents of PM2.5 is lacking. In this paper we assemble a very large data set, where we link across time and space several heterogeneous data sources on environmental exposures to PM2.5, its chemical composition, health (Medicare billing claims) and potential measured confounders (e.g. socioeconomic characteristics). We develop a Bayesian spatially-varying (SV) mortality risks model to estimate the association between long term PM2.5 and mortality and also to investigated as whether long-term exposure to some of the PM2.5 components further exacerbate the risk. A major challenge is that the number of monitor measuring the component of PM2.5 is not aligned with the monitors measuring PM2.5 total mass. Therefore we also develop a spatial modelling approach for imputing missing levels of the chemical components of PM2.5 at the monitoring locations of PM2.5 and Bayesian approaches to propagate the missing data uncertainty into the health effect estimation. Our results further advance our understanding of the toxicity of PM2.5 and can be used to generate more refined hypothesis.

email: fdominic@hsph.harvard.edu

SPATIAL SURVEILLANCE FOR NEGLECTED TROPICAL DISEASES

Lance A. Waller*, Emory University
Shannon McClintock, Emory University
Ellen Whitney, Emory University

Neglected tropical diseases (NTDs) represent a class of diseases with pockets of high prevalence and severe local impact but receive little attention compared to well-known diseases with higher global prevalence such as malaria and cholera. The focality of disease incidence and prevalence complicates establishment of accurate surveillance to provide local and global health agencies accurate information for allocation of treatment and prevention resources. Based on collaborations with the World Health Organization and the Ministry of Health in Ghana, we review data sources, known and suspected local risk factors, and analyze existing surveillance data to identify areas endemic, at risk, and currently free from Buruli ulcer. We provide a thorough assessment of the strengths and limitations of current data with respect to the identification of temporal and spatial variations in risk.

email: lwaller@emory.edu

MULTIVARIATE SPATIAL-TEMPORAL MODEL FOR BIRTH DEFECTS AND AMBIENT AIR POLLUTION RISK ASSESSMENT

Montse Fuentes*, North Carolina State University
Josh Warren, University of North Carolina, Chapel Hill
Amy Herring, University of North Carolina, Chapel Hill
Peter Langois, Texas State Health Department

We introduce a Bayesian spatial-temporal hierarchical multivariate probit regression model that identifies weeks during the first trimester of pregnancy which are impactful in terms of cardiac congenital anomaly development. The model is able to consider multiple pollutants and a multivariate cardiac anomaly grouping outcome jointly while allowing the critical windows to vary in a continuous manner across time and space. We utilize a dataset of numerical chemical model output which contains information regarding multiple species of PM2.5. Our introduction of an innovative spatial-temporal semiparametric prior distribution for the pollution risk effects allows for greater flexibility to identify critical weeks during pregnancy which are missed when more standard models are applied. The multivariate kernel stick-breaking prior is extended to include space and time simultaneously in both the locations and the masses in order to accommodate complex data settings. Simulation study results suggest that our prior distribution has the flexibility to outperform competitor models in a number of data settings. When applied to the geo-coded Texas birth data, weeks 3, 7 and 8 of the pregnancy are identified as being impactful in terms of cardiac defect development for multiple pollutants across the spatial domain.

email: fuentes@ncsu.edu

ON DYNAMIC AREAL MODELS FOR AIR QUALITY ASSESSMENT

Sudipto Banerjee*, University of Minnesota
Harrison Quick, University of Minnesota
Bradley P. Carlin, University of Minnesota

Advances in Geographical Information Systems (GIS) have led to enormous recent burgeoning of spatial-temporal databases and associated statistical modeling. Here we depart from the rather rich literature in space-time modeling by considering the setting where space is discrete (e.g. aggregated data over regions), but time is continuous. A major objective in our application is to carry out inference on gradients of a temporal process in our dataset of monthly county level asthma hospitalization rates in the state of California, while at the same time accounting for spatial similarities of the temporal process

across neighboring counties. Use of continuous time models here allows inference at a finer resolution than at which the data are sampled. Rather than use parametric forms to model time, we opt for a more flexible stochastic process embedded within a dynamic Markov random field framework. Through the matrix-valued covariance function we can ensure that the temporal process realizations are mean square differentiable, and may thus carry out inference on temporal gradients in a posterior predictive fashion. We use this approach to evaluate temporal gradients where we are concerned with temporal changes in the residual and fitted rate curves after accounting for seasonality, spatiotemporal ozone levels, and several spatially-resolved important sociodemographic covariates.

email: baner009@umn.edu

12. BAYESIAN APPROACHES TO GENOMIC DATA INTEGRATION

DECODING FUNCTIONAL SIGNALS WITH THE ROLE MODEL

Michael A. Newton*, University of Wisconsin, Madison
Qiuling He, University of Wisconsin, Madison
Zhishi Wang, University of Wisconsin, Madison

Genome-wide gene-level data are integrated in various ways with prior biological knowledge recorded in collections of gene sets (GO/KEGG/reactome,etc). Model-based approaches to this task can overcome limitations of standard one-set-at-a-time approaches, though deploying inference continues to be challenging in routine applications. I will discuss the structure and properties of a role model and our computational solutions to posterior analysis, including MCMC and integer linear programming, that operate in the high dimensional, highly constrained discrete parameter space.

email: wiscstatman@gmail.com

AN INTEGRATIVE BAYESIAN MODELING APPROACH TO IMAGING GENETICS

Marrina Vannucci*, Rice University
Francesco C. Stingo, University of Texas MD Anderson Cancer Center
Michele Guindani, University of Texas MD Anderson Cancer Center

We present a Bayesian hierarchical modeling approach for imaging genetics. We have available data from a study on schizophrenia. Our interest lies in identifying brain regions of interest (ROIs) with discriminating activation patterns between schizophrenic and control subjects, and in relating the ROIs' activations with available genetic information from single nucleotide polymorphisms (SNPs) on the subjects. For this task we develop a hierarchical mixture model that includes several innovative characteristics: it incorporates the selection of features that discriminate the subjects into separate groups; it allows the mixture components to depend on selected covariates; it includes prior models that capture structural dependencies among features. Applied to the schizophrenia data, the model leads to the simultaneous selection of a set of discriminatory ROIs and the relevant SNPs, together with the reconstruction of the correlation structure of the selected regions.

email: marina@rice.edu

BAYESIAN GRAPHICAL MODELS FOR DIFFERENTIAL PATHWAYS

Peter Mueller*, University of Texas, Austin
Riten Mitra, University of Texas, Austin
Yuan Ji, NorthShore University Health System

Graphical models can be used to characterize the dependence structure for a set of random variables. In some applications, the form of dependence varies across different subgroups or experiments. Understanding changes in the joint distribution and dependence structure across the two subgroups is key to the desired inference. Fitting a single model for the entire data could mask the differences. Separate independent analyses, on the other hand, could reduce the effective sample size and ignore the common features. We develop a Bayesian graphical model that addresses heterogeneity and implements borrowing of strength across the two subgroups by simultaneously centering the prior towards a global network. The key feature is a hierarchical prior for graphs that borrows strength across edges, resulting in a comparison of pathways across subpopulations (differential pathways) under a unified model-based framework.

email: pmueller@math.utexas.edu

13. NEW DEVELOPMENTS IN FUNCTIONAL DATA ANALYSIS

INDEX MODELS FOR SPARSELY SAMPLED FUNCTIONAL DATA

Gareth James*, University of Southern California
Xinghao Qiao, University of Southern California
Peter Radchenko, University of Southern California

The regression problem involving one or more functional predictors has many important applications. A number of methods have been developed for performing functional regression. However, a common complication in functional data analysis is one of sparsely observed curves, that is predictors that are observed, with error, on a small subset of the possible time points. Such sparsely observed data induces an errors-in-variables model where one must account for measurement error in the functional predictors. We propose a new functional errors-in-variables approach, Sparse Index Model Functional Estimation (SIMFE), which uses a functional multi index model formulation to deal with sparsely observed predictors. SIMFE has several advantages over more traditional methods. First, the multi index model allows it to produce nonlinear regressions and use an accurate supervised method to estimate the lower dimensional space into which the predictors should be projected. Second, SIMFE can be applied to both scalar and functional responses and multiple predictors. Finally, SIMFE uses a mixed effects model to effectively deal with very sparsely observed functional predictors and to correctly model the measurement error. We illustrate SIMFE on both simulated and real world data and show that it can produce superior results relative to more traditional approaches.

email: gareth@usc.edu

FUNCTIONAL DATA ANALYSIS OF GENERALIZED QUANTILE REGRESSIONS

Mengmeng Guo, Southwestern University of Finance and Economics, China
Lan Zhou*, Texas A&M University
Jianhua Huang, Texas A&M University
Wolfgang Haerdle, Humboldt University, Berlin

Generalized quantile regressions, including the conditional quantiles and expectiles as special cases, are useful alternatives to the conditional means for characterizing a conditional distribution, especially when the interest lies in the tails. We develop a functional data analysis approach to jointly estimate a family of generalized quantile regressions. Our approach assumes that the generalized quantile regressions share some common features that can be summarized by a small number of principal component functions. The principal component functions are modeled as splines and are estimated by minimizing a penalized asymmetric loss measure. An iterative least asymmetrically weighted squares algorithm is developed for computation. While separate estimation of individual generalized quantile regressions usually suffers from large variability due to lack of sufficient data, by borrowing strength across data sets, our joint estimation approach significantly improves the estimation efficiency, which is demonstrated in a simulation study. The proposed method is applied to data from 150 weather stations in China to obtain the generalized quantile curves of the volatility of the temperature at these stations.

email: lzhou@stat.tamu.edu

A GENERAL ASYMPTOIC FRAMEWORK FOR PCA CONSISTENCY

Haipeng Shen*, University of North Carolina, Chapel Hill
Dan Shen, University of North Carolina, Chapel Hill
J. S. Marron, University of North Carolina, Chapel Hill

A general asymptotic framework is developed for studying consistency properties of principal component analysis (PCA). Our framework includes several previously studied domains of asymptotics as special cases and allows one to investigate interesting connections and transitions among the various domains. More importantly, it enables us to investigate asymptotic scenarios that have not been considered before, and gain new insights into the consistency, subspace consistency and strong inconsistency regions of PCA and the boundaries among them. We also establish the corresponding convergence rate within each region. Under general spike covariance models, the dimension (or the number of variables) discourages the consistency of PCA, while the sample size and spike information (the relative size of the population eigenvalues) encourages PCA consistency. Our framework nicely illustrates the relationship among these three types of information in terms of dimension, sample size and spike size, and rigorously characterizes how their relationships affect PCA consistency.

email: haipeng@email.unc.edu

DIMENSION REDUCTION FOR SPARSE FUNCTIONAL DATA

Fang Yao*, University of Toronto
Edwin Lei, University of Toronto

In this work we propose a nonparametric dimension reduction approach for sparse functional data, aiming for enhancing the predictivity of functional regression models. In existing work of dimension reduction for functional data, there is a fundamental challenge when the functional trajectories are not fully observed or densely recorded. The goal of our work is to pool together information from sparsely observed functional objects to produce consistent estimation of the effective dimensional reduction (EDR) space. A new framework for determining the EDR space is defined so that one can borrow strength from the entire sample to recover the semi-positive definite operator that spans the EDR space. The validity of this determining operator and the asymptotic property of the resulting estimator are established. The numerical performance of the proposed method is illustrated through simulation studies and an application to an Ebay auction data.

email: fyao@utstat.toronto.edu

14. TOOLS FOR IMPLEMENTING REPRODUCIBLE RESEARCH

REPRODUCIBLE RESEARCH: SELECTING DATA OPERATIONS TOOLS

Brad H. Pollock*, University of Texas Health Science Center at San Antonio

The continuum of reproducible research begins with the process of data collection. For human studies, biostatisticians play a pivotal role in helping to develop study hypotheses, delineating required data elements and organization often through metadata, implementing quality control and data validation routines, and accessing study data for accrual/safety/efficacy monitoring and statistical analyses. If appropriate data are not collected, recorded, organized, and accessed in a comprehensive and reliable manner, data analyses will not produce valid results and inferences may be incorrect. To ensure high quality and efficiency, there are a number of important considerations which dictate the choice of tools for data operations, including: complexity of data collection requirements and validation, data organization (flat file with fixed events vs. transaction-oriented database structure), the need for curation and interoperability with other data systems (e.g., biorepositories and high-throughput omics data stores), query capability, and data access/export features. A discussion of data operations tool characteristics will be presented including: the use of electronic data capture vs. database management systems; open source, restricted-use vs. commercial software; and statistical software linkage/export capabilities.

email: bpollock@uthscsa.edu

REPRODUCIBLE RESEARCH TOOLS FOR CREATING BOOKS

Max Kuhn*, Pfizer Global R&D

Here, we will present a summary of some useful tools and workflows for creating large, multi-chapter works. The goal is to facilitate computationally demanding tasks in a large work with multiple authors while maintaining complete reproducibility for readers.

email: Max.Kuhn@pfizer.com

KNITR: A GENERAL-PURPOSE TOOL FOR DYNAMIC REPORT GENERATION IN R

Yihui Xie*, Iowa State University

Reproducible research is often related to literate programming, a paradigm conceived by Donald Knuth to combine computer code and documentation together. However, early implementations like WEB and Noweb were not suitable for data analysis and report generation, which was overcome by later tools like Sweave. The new difficulty becomes the extensibility; for example, Sweave is closely tied to LaTeX and hard to extend. The knitr package was built upon the ideas of previous tools with a framework redesigned, enabling easy and fine control of many aspects of a report. In this talk, we will demonstrate how knitr works with a variety of document formats including LaTeX, HTML and Markdown, how to speed up compilation with the cache system, how we can program a report with hook functions and work with other languages such as Python, Awk and Shell scripts in the knitr framework. We will conclude with a few significant examples including student homework, data reports, blog posts and websites built with knitr. The main design philosophy of knitr is to make reproducible research easier and more enjoyable than the common practice of copying and pasting results.

email: xie@yihui.name

15. ADAPTIVE DESIGNS FOR CLINICAL TRIALS: ACADEMIA, INDUSTRY AND GOVERNMENT

BAYESIAN ADAPTIVE DESIGN AND COMMENSURATE PRIORS FOR DEVICE SURVEILLANCE

Bradley P. Carlin*, University of Minnesota
Thomas A. Murray, University of Minnesota
Theodore C. Lystig, Medtronic Inc.
Brian P. Hobbs, University of Texas MD Anderson Cancer Center

Post-market medical device surveillance studies often have important primary objectives tied to estimating a survival function at some future time with a certain amount of precision. This talk presents the details and various operating characteristics of a Bayesian adaptive design for device surveillance, as well as a method for estimating a sample size vector (determined by the maximum sample size and a pre-set number of interim looks) that will deliver the desired power. At each interim look we assess whether we expect to achieve our goals with only the current group, or whether the achievement of such goals is extremely unlikely even for the maximum sample size. We show that our Bayesian adaptive design can outperform two non-adaptive frequentist methods currently recommended by FDA guidance documents in many settings. We also investigate the robustness of our procedures to model misspecification or changes in the trial's enrollment rate, as well as the possible usefulness of 'commensurate priors' (Hobbs et al., 2011, Biometrics) that permit adaptive borrowing from historical data when available and appropriate. This last technique offers improved estimates of the entire survival curve as compared to weighted Kaplan-Meier estimates that incorporate historical information in an ad hoc way.

email: brad@biostat.umn.edu

ADAPTIVE DESIGNS AND DECISION MAKING IN PHASE 2: IT IS NOT ABOUT THE TYPE I ERROR RATE
Brenda L. Gaydos*, Eli Lilly and Company

This presentation is intentionally provocative. One of the concerns about the use of Bayesian Adaptive Designs in drug development is the difficulty in understanding the frequentist properties of the analyses. But how important is this in phase 2? A key objective in phase 2 is to efficiently reduce uncertainty in effect and to make good decisions about the future direction of development. Statistical significance does not provide sufficient information to inform decision makers on the risk of moving into phase 3, and can lead to sub-optimal decisions. In this presentation, an approach will be presented on designing and interpreting a phase 2 Bayesian adaptive design.

email: blg@lilly.com

ADAPTIVE DESIGNS: A CBER STATISTICAL PERSPECTIVE
Estelle Russek-Cohen*, U.S. Food and Drug Administration Center for Biologics
Min A. Lin, U.S. Food and Drug Administration Center for Biologics
John A. Scott, U.S. Food and Drug Administration Center for Biologics

This past year we conducted a systematic survey of INDs that incorporated some form of adaptation in the Center for Biologics Evaluation and Research, US Food and Drug Administration. These ranged from phase 1 to phase 4 studies. This talk will capture what the range of adaptations we have seen and some of the comments expressed by the statistical reviewers have been. The use of adaptive designs has varied considerably by product area and I plan on touching on how the products differ and what appears to drive the use of adaptive designs. Some of the challenges will be also touched upon including estimation of treatment effect and analysis of secondary endpoints.

email: Estelle.Russek-Cohen@fda.hhs.gov

16. COPULAS: THEORY AND APPLICATIONS

BAYESIAN INFERENCE FOR CONDITIONAL COPULA MODELS WITH CONTINUOUS AND DISCRETE RANDOM VARIABLES
Radu V. Craiu*, University of Toronto
Avideh Sabeti, University of Toronto

The conditional copula device has opened a new range of possibilities for modelling dependence between random variables. In particular, it allows the statistician to use copulas in regression settings. Within this framework model dependence between continuous and discrete random variables in the presence of covariates. We consider a semiparametric model in which the copula parameter varies with covariates and the relationship is approximated using polynomial splines. The Bayesian paradigm considered allows simultaneous inference for the marginals and the copula. We discuss computation and model selection techniques and demonstrate the performance of the method via simulations and real data.

email: craiu@utstat.toronto.edu

TESTING HYPOTHESES FOR THE COPULA OF DYNAMIC MODELS
Bruno Remillard*, HEC Montreal

The asymptotic behavior of the empirical copula constructed from residuals of stochastic volatility models is studied. It is shown that if the stochastic volatility matrix is diagonal, then the empirical copula process behaves like if the parameters were known, a remarkable property. However, this is not true in general. Applications for goodness-of-fit and detection of structural change in the copula of the innovations are discussed.

email: bruno.remillard@hec.ca

GEOCOPULA MODELS FOR SPATIAL-CLUSTERED DATA ANALYSIS
Peter X.K. Song*, University of Michigan
Yun Bai, Fifth Third Bank

Spatial-clustered data refer to high-dimensional correlated measurements collected from units or subjects that are spatially clustered. Such data arise frequently from studies in social and health sciences. We propose a unified modeling framework, termed as GeoCopula, to characterize both large-scale variation and small-scale variation for various data types, including continuous data, binary data and count data as special cases. To overcome challenges in the estimation and inference for the model parameters, we propose an efficient composite likelihood approach in that the estimation efficiency is resulted from a construction of over-identified joint composite estimating equations. Consequently the statistical theory for the proposed estimation is developed by extending the classical theory of the generalized methods of moments. A clear advantage of the proposed estimation method is the computation feasibility. We conduct several simulation studies to assess the performance of the proposed models and estimation methods for both Gaussian and binary spatial-clustered data. Results show a clear improvement on estimation efficiency over the conventional composite likelihood method. An illustrative data example is included to motivate and demonstrate the proposed method.

email: pxsong@umich.edu

17. STOCHASTIC MODELING AND INFERENCE FOR DISEASE DYNAMICS

GRAPHICAL MODELS OF THE EFFECT HIGHLY INFECTIOUS DISEASE
Clyde F. Martin*, Texas Tech University

Graphical models for measles and influenza have been developed and studied extensively. Diseases of livestock have not been studied as extensively. In this talk we will examine a two species epidemic of hoof and mouth disease. In Texas there is a large population of feral swine and of course a major industry in beef cattle. Swine are susceptible to foot and mouth disease and tend to amplify the infection. They also move over large areas and thus present an ideal vector for spreading the disease. An outbreak of hoof and mouth disease would be devastating to the cattle industry as it would cause a total quarantine of livestock products. In this presentation we will develop a model for the spread of the disease in the two interacting populations.

email: clyde.f.martin@ttu.edu

NEW METHODS FOR ESTIMATING AND PROJECTING THE NATIONAL HIV/AIDS PREVALENCE
Le Bao*, The Pennsylvania State University

Objectives: As the global HIV pandemic enters its fourth decade, countries have collected longer time series of surveillance data, and the AIDS-specific mortality has been substantially reduced by the increasing availability of antiretroviral treatment. A refined model with a greater flexibility to fit longer time series of surveillance data is desired. Methods: In this article, we present a new epidemiological model that allows the HIV infection rate, r(t), to change over years. The annual change of infection rate is modeled by a linear combination of three key factors: the past prevalence, the past infection rate, and a stabilization condition. We focus on fitting the antenatal clinic (ANC) data and household surveys which are the most commonly available data source for generalised epidemics defined by the overall prevalence being above 1%. Results and Conclusion: The proposed model better captures the main pattern of the HIV/AIDS dynamic. The three factors in the proposed model all have significant contributions to the reconstruction of r(t) trends. It improves the prevalence fit over the classic EPP model, and provides more realistic projections when the classic model encounters problems.

email: lebao@psu.edu

SPATIAL POINT PROCESSES AND INFECTIOUS DYNAMICS IN CELL CULTURES

John Fricks*, The Pennsylvania State University

The interaction of multiple viral strains in a single organism is important to both the infection process and the immunological response to infection. An in vitro cell culture system is studied using spatial point process to analyze the interaction of multiple viral strains of Newcastle Disease Virus (NDV) measured through fluorescent markers. Both exploratory tools and mechanistic models of interaction will be discussed.

email: fricks@stat.psu.edu

18. MODEL SELECTION FOR HIGH-DIMENSIONAL GENETICS DATA

REGULARIZED INTEGRATIVE ANALYSIS OF CANCER PROGNOSIS STUDIES

Jin Liu*, Yale University
Jian Huang, University of Iowa
Shuangge Ma, Yale University

In cancer prognosis studies, an essential goal is to identify a small number of genetic markers that are associated with disease-free or overall survival. A series of recent studies have shown that integrative analysis, which simultaneously analyzes multiple datasets, is more effective than the analysis of single datasets and meta-analysis. With multiple prognosis datasets, their genomic basis can be characterized using the homogeneity model or the heterogeneity model. The heterogeneity model allows overlapping but possibly different sets of markers for different datasets, includes the homogeneity model as a special case, and can be more flexible. In this study, we analyze multiple heterogeneous cancer prognosis datasets and adopt the AFT (accelerated failure time) model to describe survival. A weighted least squares approach is adopted for estimation. For marker selection, we propose using sparse group penalization approaches, in particular, sparse group Lasso (SGLasso) and sparse group MCP (SGMCP). The proposed penalties are the sum of a group penalty and a penalty on individual marker. A group coordinate descent approach is developed to compute the proposed estimates. Simulation study shows satisfactory performance of the proposed approaches. We analyze three lung cancer prognosis datasets with microarray gene expression measurements using the proposed approaches.

e-mail: jin.liu.jl2329@yale.edu

A BAYESIAN DIMENSION REDUCTION APPROACH FOR DETECTION OF MULTI-LOCUS INTERACTION IN CASE-CONTROL STUDIES

Debashree Ray*, University of Minnesota
Xiang Li, University of Minnesota
Wei Pan, University of Minnesota
Saonli Basu, University of Minnesota

Genome-wide association studies (GWAS) provide a powerful approach for detection of single nucleotide polymorphisms (SNPs) associated with complex diseases. Single-locus association analysis is a primary tool in GWAS but it has low power in detecting SNPs with small effect sizes and in capturing gene-gene interaction. Multi-locus association analysis can improve power to detect such interactions by jointly modelling the SNP effects within a gene and by reducing the burden of multiple hypothesis testing in GWAS, but such multivariate tests have large degrees of freedom that can also compromise power. We have proposed here a powerful and flexible dimension reduction approach to detect multi-locus interaction. We use a Bayesian partitioning model which clusters SNPs according to their direction of association, models higher order interactions using a flexible scoring scheme, and uses a model-averaging approach to detect association between the SNP-set and the disease. For any SNP-set, only three parameters are needed to model multi-locus interaction, which is considerably less than any other competitive parametric method. We have illustrated our model through extensive simulation studies and have shown that our approach has better power than some of the currently popular multi-locus approaches in detecting genetic variants associated with a disease under various epistatic models.

e-mail: rayxx267@umn.edu

GENE-GENE AND GENE-ENVIRONMENT INTERACTIONS: BEYOND THE TRADITIONAL LINEAR MODELS

Yuehua Cui*, Michigan State University

The genetic architecture of complex diseases involves complicated gene-gene (G×G) and gene-environment (G×E) interactions. Identifying G×G and G×E interactions has been the major theme in genetic association studies where linear models have been broadly applied. In this talk, I will present some of our recent work in this direction. For the study of G×G interaction, I will illustrate how to model the joint effect of multiple variants in a gene using a kernel machine method to detect the interaction from a gene level rather than from a single SNP level. For the G×E interaction, we relax the commonly assumed linear interaction assumption and allow non-linear genetic penetrance under different environmental stimuli to identify which genes and in what format they respond to environment changes. Application to real data will be given to show the utility of the work.

e-mail: cui@stt.msu.edu

A VARIABLE-SELECTION-BASED NOVEL STATISTICAL APPROACH TO IDENTIFY SUSCEPTIBLE RARE VARIANTS ASSOCIATED WITH COMPLEX DISEASES WITH DEEP SEQUENCING DATA

Hokeun Sun*, Columbia University
Shuang Wang, Columbia University

Existing association methods on sequencing data have been focused on aggregating variants across a gene or a genetic region in the past years due to the fact that analyzing individual rare variants is underpowered. However, to identify which rare variants in a gene or a genetic region out of all variants are associated with the outcomes (either quantitative or qualitative) is a natural next step. We proposed a variable-selection-based novel approach that is able to identify the locations of the susceptible rare variants that are associated with the outcomes with sequencing data. Specifically, we generated the power set of the p rare variants except the empty set. We then treated the K=2^p-1 subsets as the K 'new variables' and applied the penalized likelihood estimation using L1-norm regularization. After selecting the most associated subset with the outcome, we applied a permutation procedure specifically designed to assess the statistical significance of the selected subset. In simulation studies, we demonstrated that the proposed method is able to select subsets with most of the outcome related rare variants. The type I error and power of the subsequent permutation procedure demonstrated the validity and advantage of our selection method. The proposed method was also applied to sequence data on the ANGPTL family of genes from the Dallas Heart Study (DHS).

e-mail: hs2674@columbia.edu

A NOVEL METHOD TO CORRECT PARTIALLY SEQUENCED DATA FOR RARE VARIANT ASSOCIATION TEST

Song Yan*, University of North Carolina, Chapel Hill

Despite its great capacity to detect rare-variant and complex-trait associations, the cost of next-generation sequencing is still very high for data with a large number of individuals. In the case and control studies, it is thus appealing to sequence only case individuals and genotype the remaining ones since the causal SNPs are usually enriched in the case group. However, it is well known that this approach leads to inflated type-I error estimation. Several methods have been proposed in the literature to correct the type-I error bias but all underpowered. We propose a novel method which not only corrects the bias but also achieves a higher testing power compared with the existing methods.

e-mail: songyan@unc.edu

ChIP-Seq OUT, ChIP-exo IN?
Dongjun Chung*, Yale University
Irene Ong, University of Wisconsin, Madison
Jeffrey Grass, University of Wisconsin, Madison
Robert Landick, University of Wisconsin, Madison
Sunduz Keles, University of Wisconsin, Madison

ChIP-exo is a recently proposed modified ChIP-Seq protocol that aims to attain high resolution in the identification of protein-DNA interaction sites by using exonuclease. In spite of its great potential to improve resolution compared to the conventional ChIP-Seq experiments, the current literature lack methods that fully take advantage of ChIP-exo data. In order to address this, we developed dPeak, an algorithm that analyzes both ChIP-exo and ChIP-Seq data in a unified framework. The dPeak algorithm implements a probabilistic model that accurately describes each of ChIP-exo, PET ChIP-Seq, and SET ChIP-Seq data generation processes. In order to evaluate benefits of ChIP-exo rigorously, we generated both ChIP-exo and ChIP-Seq data for sigma70 factor in Escherichia coli, which requires distinction of closely spaced binding events separated by only few base pairs. Using the dPeak algorithm and the sigma70 data, we rigorously compared ChIP-exo and ChIP-Seq from resolution point of view and investigated design issues regarding ChIP-exo. Our results have important implications for the design of ChIP-Seq and ChIP-exo experiments.

e-mail: dongjun.chung@yale.edu

FUNCTIONAL MIXED EFFECTS MODELS FOR IMAGING GENETIC DATA
Ja-An Lin*, University of North Carolina, Chapel Hill
Hongtu Zhu, University of North Carolina, Chapel Hill
Wei Sun, University of North Carolina, Chapel Hill
Jiaping Wang, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

Traditional statistical methods analyzing imaging genetics data, which includes medical imaging measurements (e.g. MRI) and genetic variation (e.g. SNP), often suffer from low statistical power. We propose a method named functional mixed effects model (FMEM) to address this issue. It incorporates a group of genetic variation as a population- shared random effect with a common variance component (VC), and other clinical variables as fixed effect in a weighted likelihood model. Then we adaptively include neighbourhood voxels with certain weights while estimating both the VC and the regression coefficients. The integrated genomic effect is investigated by testing the zero of the VC via weighted likelihood ratio test with similar adaptive strategy involving nearby voxels. The performance of FMEM is evaluated by simulation studies and the result shows it outperforms voxel-wise based approach by greater statistical power with reasonable type I error control. FMEM is also applied to the ADNI study to identify the brain regions affected by the candidate gene TOMM40 in patients with Alzheimer's disease that the FMEM finds 28 regions while the classical voxel-wise approach only finds 6.

e-mail: jaanlin@live.unc.edu

19. CAUSAL INFERENCE

CAUSAL INFERENCE FOR THE NONPARAMETRIC MANN-WHITNEY-WILCOXON RANK SUM TEST
Pan Wu*, University of Rochester

The Mann-Whitney-Wilcoxon (MWW) rank sum test is widely used for comparing two treatment groups, especially when there are outliers in the data. However, MWW generally yields invalid conclusions when applied to non-randomized studies, particularly those in epidemiologic research. Although one may control for selection bias by including covariates in regression analysis or use available formal causal methods such as the Propensity Score and Marginal Structural Models, such analyses yield results that are not only subjective based on how the outliers are handled, but also are often difficult to interpret. Rank based methods such as the MWW test are more effective to address such extremely large outcomes. In this paper, we extend the MWW rank sum test to provide causal inference for non-randomized study data by integrating the counterfactual outcome paradigm with the functional response models (FRM), which is uniquely positioned to model dynamic relationships between subjects, rather than attributes of a single subject as in most regression models, such as the MWW test within our context. The proposed approach is illustrated with data from both real and simulated studies.

email: pan_wu@urmc.rochester.edu

SHARPENING BOUNDS ON PRINCIPAL EFFECTS WITH COVARIATES
Dustin M. Long*, West Virginia University
Michael G. Hudgens, University of North Carolina, Chapel Hill

Estimation of treatment effects in randomized studies is often hampered by possible selection bias induced by conditioning on or adjusting for a variable measured post-randomization. One approach to obviate such selection bias is to consider inference about treatment effects within principal strata, i.e., principal effects. A challenge with this approach is that without strong assumptions principal effects are not identifiable from the observable data. In settings where such assumptions are dubious, identifiable large sample bounds may be the preferred target of inference. In practice these bounds may be wide and not particularly informative. In this work we consider whether bounds on principal effects can be improved by adjusting for a categorical baseline covariate. Adjusted bounds are considered which are shown to never be wider than the unadjusted bounds. Necessary and sufficient conditions are given for which the adjusted bounds will be sharper (i.e., narrower) than the unadjusted bounds. The methods are illustrated using data from a recent, large study of interventions to prevent mother-to-child transmission of HIV through breastfeeding. Using a baseline covariate indicating low birth weight, the estimated adjusted bounds for the principal effect of interest are 64% narrower than the estimated unadjusted bounds.

email: dmlong@hsc.wvu.edu

MODEL AVERAGED DOUBLE ROBUST ESTIMATION
Matthew Cefalu*, Harvard School of Public Health
Francesca Dominici, Harvard School of Public Health
Giovanni Parmigiani, Dana-Farber Cancer Institute and Harvard School of Public Health

Existing methods in causal inference do not account for the uncertainty in the selection of confounders. We propose a new class of estimators for the average causal effect, the model averaged double robust estimators, that formally account for model uncertainty in both the propensity score and outcome model through the use of Bayesian model averaging. These estimators build on the desirable double robustness property by only requiring the true propensity score model or the true outcome model be within a specified class of models to maintain consistency. We provide asymptotic results and conduct a large scale simulation study that indicates the model averaged double robust estimator has better finite sample behavior than the usual double robust estimator.

email: mcefalu@fas.harvard.edu

NEW APPROACHES FOR ESTIMATING PARAMETERS OF STRUCTURAL NESTED MODELS
Edward H. Kennedy*, University of Pennsylvania School of Medicine
Marshall M. Joffe, University of Pennsylvania School of Medicine

Structural nested models (SNMs) are a powerful way to represent causal effects in longitudinal studies with treatments that change over time. They can handle time-dependent effect modification and instrumental variables, for example, while appropriately controlling for confounding and without requiring distributional assumptions. However, SNMs are used very rarely in real applications. This is at least partially due to practical difficulties associated with semiparametric g-estimation, the standard approach for estimation of SNM parameters. In this work we explore modifications of and alternative approaches to g-estimation for SNMs, including parametric likelihood-based inference, generalized method of mo-

ments, and empirical likelihood. We develop frameworks for each approach, establish the asymptotic properties of corresponding estimators, and investigate finite sample performance via comprehensive simulation studies. We also consider relaxing standard ignorability assumptions by allowing for the presence of unmeasured confounding in subsets of the data. We apply our methods using observational data to examine the effect of erythropoietin on mortality for Medicare patients on hemodialysis.

email: edwardh.kennedy@gmail.com

MARGINAL STRUCTURAL COX MODELS WITH CASE-COHORT SAMPLING

Hana Lee*, University of North Carolina, Chapel Hill
Michael G. Hudgens, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill

An objective of biomedical cohort studies often entails assessing the effect of a time-varying treatment or exposure on a survival time. In the presence of time-varying confounders, marginal structural models fit using inverse probability weighting can be employed to obtain a consistent and asymptotically normal estimator of the causal effect of a time-varying treatment. This article considers estimation of parameters in the semiparametric marginal structural Cox model (MSCM) from a case-cohort study. Case-cohort sampling entails assembling covariate histories only for cases and a random subcohort, which can be cost effective, particularly in large cohort studies with low incidence rates. Following Cole et al. (2012), we consider estimating the causal hazard ratio from a MSCM by maximizing a weighted-pseudo-partial-likelihood. The estimator is shown to be consistent and asymptotically normal under certain regularity assumptions.

email: hanalee@email.unc.edu

TARGETED MINIMUM LOSS-BASED ESTIMATION OF A CAUSAL EFFECT ON AN OUTCOME WITH KNOWN CONDITIONAL BOUNDS

Susan Gruber*, Harvard School of Public Health
Mark J. van der Laan, University of California, Berkeley

Targeted minimum loss based estimation (TMLE) provides a framework for locally efficient double robust causal effect estimation. This talk describes a TMLE that incorporates known conditional bounds on a continuous outcome. Subject matter knowledge regarding the bounds of a continuous outcome within strata defined by a subset of covariates, X, translates into statistical knowledge that constrains the model space of the true joint distribution of the data. In settings where there is low Fisher Information in the data for estimating the desired parameter, as is common when X is high dimensional relative to sample size, incorporating this domain knowledge can improve the fit of the targeted outcome regression, thereby improving bias and variance of the parameter estimate. TMLE, a substitution estimator defined as a mapping from a density to a (possibly d-dimensional) real number, readily incorporates this global knowledge, resulting in improved finite sample performance.

email: sgruber@hsph.harvard.edu

SURROGACY ASSESSMENT USING PRINCIPAL STRATIFICATION WHEN SURROGATE AND OUTCOME MEASURES ARE MULTIVARIATE NORMAL

Anna SC Conlon*, University of Michigan
Jeremy MG Taylor, University of Michigan
Michael R. Elliott, University of Michigan

In clinical trials, a surrogate outcome variable (S) can be measured before the outcome of interest (T) and may provide early information regarding the treatment (Z) effect on T. Most previous methods for surrogate validation rely on models for the conditional distribution of T given Z and S. However, S is a post-randomization variable, and unobserved, simultaneous predictors of S and T may exist, resulting in a noncausal interpretation. Using the principal surrogacy framework introduced by Frangakis and Rubin (2002), we propose a Bayesian estimation strategy for surrogate validation when the joint distribution of potential surrogate and outcome measures is multivariate normal. We model the joint conditional distribution of the potential outcomes of T, given the potential outcomes of S and propose surrogacy validation measures from this model. By conditioning on principal strata of S, the resulting estimates are causal. As the model is not fully identifiable from the data, we propose some reasonable prior distributions and assumptions that can be placed on weakly identified parameters to aid in estimation. We explore the relationship between our surrogacy measures and the traditional surrogacy measures proposed by Prentice (1989). The method is applied to data from a macular degeneration study and data from an ovarian cancer study, both previously analyzed by Buyse, et al. (2000).

email: achern@umich.edu

20. HEALTH SERVICES AND HEALTH POLICY RESEARCH

RISK-ADJUSTED INDICES OF COMMUNITY NEED USING SPATIAL GLMMs

Glen D. Johnson*, Lehman College, CUNY School of Public Health

Funding allocation for community-based public health programs is often not based on objective, evidence-based, decision-making. A quantitative assessment of actual community need would provide a means for rank-ordering and prioritizing communities. Past indices of community need or deprivation are applications of multivariate methods for reducing a large set of variables to a generic index; however, public health programs are driven by particular outcomes such as the caseload of teen pregnancies, diabetes, obesity, etc. A solution for estimating the caseload (or rate) by community, defined by some sub-county geographic delineation, in a way that also incorporates other community variables, is to apply generalized linear models with a random effect for residual spatial autocorrelation. This approach has been successfully applied for estimating teen pregnancy and sexually transmitted disease incidence by postal ZIP code in New York State where it has been used for guiding adolescent pregnancy prevention programs. This case study will be demonstrated as an application of negative binomial regression with a discussion of various modeling approaches, from a simple spatial random addition to the intercept, solved through pseudo-likelihood, to a CAR model that is solved through Bayesian MCMC methods. Approaches to geo-visualizing results will also be shared.

email: glen.johnson@lehman.cuny.edu

DATA ENHANCEMENTS AND MODELING EFFORTS TO INFORM RECENT HEALTH POLICY INITIATIVES

Steven B. Cohen*, Agency for Healthcare Research and Quality

Existing sentinel health care databases that provide nationally representative population based data on measures of health care access, cost, use, health insurance coverage, health status and health care quality, provide the necessary foundation to support descriptive and behavioral analyses of the U.S. health care system. Given the recent passage of the Affordable Care Act (ACA) and the rapid pace of changes in the financing and delivery of health care, policymakers, health care leaders and decision makers at both the national and state level are particularly sensitive to recent trends in health care costs, coverage, use, access and health care quality. Government and non-governmental entities rely upon these data to evaluate health reform policies, the effect of tax code changes on health expenditures and tax revenue, and proposed changes in government health programs. AHRQ's Medical Expenditure Panel Survey is one of the

core data resources utilized to inform several provisions of the Affordable Care Act. In this presentation, attention will be given to the current capacity of the MEPS and survey enhancements to inform program planning, implementation, and evaluations of program performance for several components of the ACA. The presentation will also highlight recent research findings and modeling efforts to inform ongoing health reform initiatives.

email: scohen@ahrq.gov

ESTIMATE THE TRANSITION PROBABILITY OF DISEASE STAGES USING LARGE HEALTHCARE DATABASES WITH A HIDDEN MARKOV MODEL

Lola Luo*, University of Pennsylvania
Dylan Small, University of Pennsylvania
Jason A. Roy, University of Pennsylvania

Chronic kidney disease (CKD) is a world-wide public health problem and according to the National Kidney Foundation, 26 million American adults have CKD and millions of others are at increased risk. Since CKD is incurable, information about the disease progression is relevant to the researchers. One way to learn more about CKD is from large healthcare databases. They are easy to obtain, contain a lot of information, and naturalistic as oppose to the controlled such as in clinical trials. We have obtained such data from Geisinger Health Clinics in Pennsylvania. Since the data is not from a controlled study, the characteristic of data can get fairly complex due to disorganized visits and measurement errors, especially when data are large. This will often cause problems for making inference on the data. We have proposed a discretization method to transform a continuous time Markov process to a discrete time Markov process. Then, we used a discrete time hidden Markov model with five hidden states and five observable states to estimate the transition probability and the measurement error of the CKD data.

email: luolola@mail.med.upenn.edu

MEDIAN COST ASSOCIATED WITH REPEATED HOSPITALIZATIONS IN PRESENCE OF TERMINAL EVENT

Rajeshwari Sundaram*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Subhshis Ghoshal, North Carolina State University
Alexander C. McLain, University of South Carolina

Recurrent events data are often encountered in longitudinal follow-up studies. Often in such studies, the observation of recurrent events gets terminated by informative dropouts or terminal events. Example of such data is repeated hospitalization in HIV-positive patients or cancer patients. In such situations, the occurrence of terminal events precludes from further recurrences. Our motivation is the cost associated with treating such recurrences. We propose a two-stage modeling approach where the costs are modeled flexibly using a parametric model and a joint model for the underlying recurrent event process and terminal event using a shared frailty to capture dependence between the recurrent event process and time to terminal event. The model for the recurrent event process is a non-stationary Poisson process with covariate effects multiplicative on baseline intensity function and the time to terminal event is a proportional hazards model, conditional on the frailty. We propose an estimating equation approach to estimating the median cost associated with the recurrent event process based on estimates of the underlying models for recurrent events and time to terminal event as proposed in Huang and Wang (2004). Our proposed estimation procedure is investigated through simulation studies as well as applied to SEER-Medicare data on medical cost associated with ovarian cancer.

email: sundaramr2@mail.nih.gov

DOUBLY ROBUST ESTIMATION OF THE COST-EFFECTIVENESS OF REVASCULARIZATION STRATEGIES

Zugui Zhang*, Christiana Care Health System
Paul Kolm, Christiana Care Health System
William S. Weintraub, Christiana Care Health System

Selection bias has been one of the important issues in observational studies when the controls are not representative of the study base, and estimation of the effect of a treatment with a causal interpretation from observational studies could be misleading due to confounding. The double robust estimator combines both inverse propensity score weighting and Regression modeling and therefore offers protection against mis-modeling in observational studies. The purpose of this study was to apply doubly Robust Estimation to examine the cost-effectiveness of coronary-artery bypass grafting (CABG) versus percutaneous coronary intervention (PCI), using data from the Society of Thoracic Surgeons (STS) Database and the American College of Cardiology Foundation (ACCF) National Cardiovascular Data Registry (NCDR) in ASCERT. The STS Database and ACCF NCDR were linked to the Centers for Medicare and Medicaid Services (CMS) claims data from years 2004 to 2008. Costs were assessed at index, 30days, 1 year and follow-up years by Diagnosis Related Group for hospitalizations. Effectiveness was measured via mortality rate, incidence of stroke, and actual MI and converted to life year gained (LYG) from Framingham survival data. Costs and effectiveness were adjusted using doubly Robust Estimation via propensity scores and inverse probability weighting to reduce treatment selection bias.

email: zzhang@christianacare.org

COMPUTING STANDARDIZED READMISSION RATIOS BASED ON A LARGE SCALE DATA SET FOR KIDNEY DIALYSIS FACILITIES WITH OR WITHOUT ADJUSTMENT OF HOSPITAL EFFECTS

Jack D. Kalbfleisch, University of Michigan
Yi Li, University of Michigan
Kevin He*, University of Michigan
Yijiang Li, Google

The purpose of this study is to evaluate the impact of including discharging hospitals on the estimation of Facility-level Standardized Readmission Ratios (SRRs). The motivating example is the evaluation of readmission rates among kidney dialysis facilities in the United States. The estimation of SRRs consists of two steps. First, we model the dependence of readmission events on facilities and patient-level characteristics, with or without the adjustment for discharging hospitals. Second, we use results from the model to compute the SRR for a given facility as a ratio of the number of observed events to the number of expected events given the case-mix in that facility. A challenge part in our motivating example is that the number of parameters is very large and estimation of high-dimensional parameters is troublesome. To solve this problem, we propose a Newton-Raphson algorithm for logistic fixed effect model and an approximate EM algorithmn for the logistic mixed effect model. We also considered a re-sampling and simulation technique to derive P-values for the proposed measures. The finite-sample properties of proposed measures are examined through simulation studies. The methods developed are applied to national kidney transplant data.

email: kevinhe@umich.edu

MULTIPLE MEDIATION IN CLUSTER-RANDOMISED TRIALS

Sharon Wolf, New York University
Elizabeth L. Turner*, Duke University
Margaret Dubeck, College of Charleston
Simon Brooker, London School of Hygiene and Tropical Medicine and Kenyan Medical Research Institute
Matthew Jukes, Harvard University

Cluster-randomized trials (CRTs) are often used to test the effect of multi-component interventions. It is of scientific interest to identify pathways (i.e. mediators) through which the intervention has an effect on outcomes. Additionally, it is hoped that in so doing the intervention can be refined to further target the identified pathways so that it can be implemented on a larger scale in a cost-effective manner. In contrast to much of the mediation literature whereby models with a single mediator are proposed, multi-component interventions typically give rise to models of multiple-mediation. Such settings present particular challenges including: defining

and identifying individual mediation effects, accounting for the clustered design of the trial and implications of mediators measured at the cluster-level on individual-level outcomes. The Health and Literacy Intervention (HALI) CRT evaluated the effects of a multi-component literacy-training program of teachers on literacy outcomes of 2500 children in 101 schools in coastal Kenya. Cluster-level mediators considered include literacy-related strategies used in the classroom. Using the HALI trial as an example, methods for the assessment of multiple-mediation are compared and contrasted, with a particular focus on bootstrapping strategies. We highlight the implications of cluster-level mediators for the analysis of individual-level outcomes.

email: liz.turner@duke.edu

21. PREDICTION / PROGNOSTIC MODELING

PREDICTING TREATMENT RESPONSE FOR RHEUMATOID ARTHRITIS PATIENTS WITH ELECTRONIC MEDICAL RECORDS

Yuanyuan Shen*, Harvard School of Public Health

Electronic medical records (EMRs) used as part of routine clinical care have great potential to serve as a rich resource of data for clinical research. One of the most challenging bottlenecks with EMR research is to precisely define patient phenotypes. Much progress has been made in recent years to develop automated algorithms for classifying patient level phenotypes by combining information from structured variables and lab results via natural language processing (NLP). However, accurate classification of phenotypes that involves temporal trajectories remains challenging due to the high dimensionality of the features that may change over time. In this study of identifying rheumatoid arthritis (RA) patients who respond to anti-TNF therapies using EMR systems in two large hospitals in Boston, we approached these problems by first developing algorithms to predict disease activity (DAS) for each patient at each hospital visit. To accommodate the high dimensionality of the features, univariate screening followed by shrinkage estimation was used to select important features. To incorporate the potential non-linear effects of the features, the final algorithm was developed based on the support vector machine using the informative features. These predicted DAS variables over time are subsequently used as functional predictors to construct an algorithm for predicting treatment response.

email: yushen@hsph.harvard.edu

A SIMPLE PLUS/MINUS METHOD FOR DISCRIMINATION IN GENOMIC DATA ANALYSIS

Sihai Zhao*, University of Pennsylvania
Levi Waldron, Harvard School of Public Health and Dana Farber Cancer Institute
Curtis Huttenhower, Harvard School of Public Health
Giovanni Parmigiani, Harvard School of Public Health and Dana Farber Cancer Institute

Recent work has shown that the discrimination power of simple classification approaches can match that of much more sophisticated procedures. The ease of implementation and interpretation of these simple approaches, however, make them more amenable to translation into the clinic. In this paper we study a simple approach we call the "plus/minus method", which calculates prognostic scores for discrimination by summing the covariates, weighted by the signs of their associations with the outcome. We show theoretically that it is a low-variability estimator that can perform almost as well as the optimal linear risk score. Simulations and an analysis of a real problem in ovarian cancer confirm that the plus/minus method can match the discrimination power of more established methods, and with a significant advantage in speed.

email: sihai@mail.med.upenn.edu

EVALUATION OF GENE SIGNATURE FOR CLINICAL ASSOCIATION BY PRINCIPAL COMPONENT ANALYSIS

Dung-Tsa Chen*, Moffitt Cancer Center
Ying-Lin Hsu, National Chung Hsing University, Taiwan

Gene expression profiling allows us to measure thousands of genes simultaneously. The technology provides better understanding about what genes promote disease development and help scientists narrow down to a small subset of genes associated with disease status when their expressions are changed. The subset of genes is often referred as the gene signature, which has unique characteristics to delineate disease status. Various developed gene signatures have provided promising means of personalized medicine. However, evaluation of a gene signature is nontrivial. One important issue is how to integrate all the genes as a whole to represent the signature in order to test its association with clinical outcomes. In this study, we will demonstrate the use of principal component analysis is feasible to evaluate a gene signature for its clinical association. Examples of various cancer datasets will be used for illustration.

email: Dung-Tsa.Chen@moffitt.org

A MODEL-FREE MACHINE LEARNING METHOD FOR SURVIVAL PROBABILITY PREDICTION

Yuan Geng*, North Carolina State University
Wenbin Lu, North Carolina State University
Hao H. Zhang, North Carolina State University

Survival probability prediction is of great interest in many medical studies since it plays an important role for patients' risk prediction and stratification. We propose a model-free machine learning method for risk group classification and survival probability prediction based on weighted support vector machine (SVM). The new method does not require to specify a parametric or semiparametric model for survival probability prediction and it can naturally handle high dimensional covariates and nonlinear covariate effects. Simulation studies are conducted to demonstrate the finite sample performance of our proposed method under various settings. Application to a glioma tumor data is also given to illustrate the methodology.

email: ygeng@ncsu.edu

SURVIVAL ANALYSIS OF CANCER DATA USING THE RANDOM FORESTS, AN ENSEMBLE OF TREES

Bong-Jin Choi*, University of South Florida
Chris P. Tsokos, University of South Florida

Tree-based methods have become popular for performing survival analysis with complex data structures such as the SEER data. Within the Random Forest (RF), we applied decision tree analysis (DTA) to identify the most important attributable variable that significantly contributes to estimating the survival time of a given cancer patient and proceed to develop a statistical model to estimate the survival time. The proposed approach is reducing the prediction error of the subject estimate using R and My-SQL. We validate the final model using the Brier score and compare the results of the developed model with the classical models.

email: bchoi@mail.usf.edu

LANDMARK ESTIMATION OF SURVIVAL AND TREATMENT EFFECT IN A RANDOMIZED CLINICAL TRIAL

Layla Parast*, RAND Corporation
Lu Tian, Stanford University
Tianxi Cai, Harvard University

In many studies with a survival outcome, it is often not feasible to fully observe the primary event of interest. This often leads to heavy censoring and thus, difficulty in efficiently estimating survival or comparing survival rates between two groups. In certain diseases, baseline covariates and the event time of non-fatal intermediate events may be associated with overall survival. In these settings,

incorporating such additional information may lead to gains in efficiency in estimation of survival and testing for a difference in survival between two groups. Most existing methods for incorporating intermediate events and covariates to predict survival focus on estimation of relative risk parameters and/or the joint distribution of events under semiparametric models. However, in practice, these model assumptions may not hold and hence may lead to biased estimates of the marginal survival. In this paper, we propose a semi-nonparametric two-stage procedure to estimate and compare t-year survival rates by incorporating intermediate event information observed before some landmark time. In a randomized clinical trial setting, we further improve efficiency through an additional augmentation step. Simulation studies demonstrate substantial potential gains in efficiency in terms of estimation and power. We illustrate our proposed procedures using an AIDS Clinical Trial dataset.

email: parast@rand.org

A BAYESIAN APPROACH TO ADAPTIVELY DETERMINING THE SAMPLE SIZE REQUIRED TO ASSURE ACCEPTABLY LOW RISK OF UNDESIRABLE ADVERSE EVENTS

A. Lawrence Gould*, Merck Research Laboratories
Xiaohua Douglas Zhang, Merck Research Laboratories

An emerging concern with new therapeutic agents, especially treatments for Type 2 diabetes, a prevalent condition that increases an individual's risk of heart attack or stroke, is the likelihood of adverse events, especially cardiovascular events that the new agents may cause. These concerns have led to regulatory requirements for demonstrating that a new agent increases the risk of an adverse event relative to a control by no more than, say, 30% or 80% with high (e.g., 97.5%) confidence. We describe a Bayesian adaptive procedure for determining if the sample size for a development program needs to be increased and, if necessary, by how much, to provide the required assurance of limited risk. The decision is based on the predictive likelihood of a sufficiently high posterior probability that the relative risk is no more than a specified bound. Allowance can be made for between-center as well as within-center variability to accommodate large-scale developmental programs, and design alternatives (e.g., many small centers, few large centers) for obtaining additional data if needed can be explored. Binomial or Poisson likelihoods can be used, and center-level covariates can be accommodate. The predictive likelihoods are explored under various conditions to assess the statistical properties of the method.

email: goulda@merck.com

22. CLUSTERING ALGORITHMS FOR BIG DATA

IDENTIFICATION OF BIOLOGICALLY RELEVANT SUBTYPES VIA PREWEIGHTED SPARSE CLUSTERING

Sheila Gaynor*, University of North Carolina, Chapel Hill
Eric Bair, University of North Carolina, Chapel Hill

When applying clustering algorithms to high-dimensional data sets, particularly DNA microarray data, the clustering results are often dominated by groups of features with high variance and high correlation with one another. Often times such clusters are of lesser interest, and we may seek to obtain clusters formed by other features with lower variance that may be more interesting biologically. In particular, in many applications one seeks to identify clusters associated with a particular outcome. We propose a method for identifying such secondary clusters using a modified version of the sparse clustering method of Witten and Tibshirani (2010). With an appropriate choice of initial weights, it is possible to identify clusters that are unrelated to previously identified clusters or clusters that are associated with an outcome variable of interest. We show that our proposed method outperforms several competing methods on simulated data sets and show that it can identify clinically relevant clusters in patients with chronic orofacial pain and clusters associated with patient survival in cancer microarray data.

email: smgaynor@live.unc.edu

BICLUSTERING VIA SPARSE CLUSTERING

Qian Liu*, University of North Carolina, Chapel Hill
Eric Bair, University of North Carolina, Chapel Hill

Biclustering methods are unsupervised learning techniques that seek to identify homogeneous groups of observations and features (i.e. checkerboard patterns) in a data set. We propose a novel biclustering method based on the sparse k-means clustering algorithm of Witten and Tibshirani. Sparse clustering performs clustering using a subset of the features by assigning a weight to each feature and giving greater weight to features that have larger between-cluster sums of squares. We demonstrate that the feature weights calculated by the sparse clustering algorithm can be used to identify biclusters in a data set. The proposed method is fast and can be easily scaled to high-dimensional data sets. We demonstrate that our proposed method produces satisfactory results in a variety of simulated data sets and apply the method to a series of cancer microarray data sets as well as data collected from the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study. Methods to evaluate the reproducibility of the biclusters identified by the algorithm are also considered.

email: qliu@live.unc.edu

BICLUSTERING WITH HETEROGENEOUS VARIANCE

Guanhua Chen*, University of North Carolina, Chapel Hill
Patrick F. Sullivan, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

In cancer research, as in all of medicine, it is important to classify patients into etiologically and therapeutically relevant subtypes to improve diagnosis and treatment. One way to do this is by using clustering methods to find subgroups of homogeneous individuals based on genetic profiles together with heuristic clinical analysis. A notable drawback of existing clustering methods is that they ignore the possibility that the variance of gene expression profile measurements can be heterogeneous across subgroups, leading to inaccurate subgroup prediction. In this paper, we present a statistical approach that can capture both mean and variance structure in gene expression data. We demonstrate the strength of our method in both synthetic data and two cancer data sets. In particular, our method confirms the hypervariability of methylation level in cancer patients, and it detects clearer subgroup patterns in lung cancer data.

email: guanhuac@live.unc.edu

BICLUSTERING WITH THE EM ALGORITHM

Prabhani Kuruppumullage Don*, The Pennsylvania State University
Bruce G. Lindsay, The Pennsylvania State University
Francesca Chiaromonte, The Pennsylvania State University

Cluster analysis is the identification of natural groups in data. Most clustering applications focus on one-way clustering, grouping observations that are similar to each other based on a set of features, or features that are similar to each other across a given set of observations. With large amount of data arising from applications such as gene expression studies, text mining studies, etc., there has been renewed interest in bi-clustering methods that group observations and features simultaneously. In one-way clustering, it is known that mixture-based techniques using the EM algorithm provide better performance, as well as an assessment of uncertainty. In bi-clustering however, evaluating the mixture likelihood using an EM algorithm is computationally infeasible and approximations are essential. In this work, we propose an approach based on a composite likelihood approximation and a nested EM algorithm to maximize the likelihood. Further, we discuss common statistical issues such as labeling and model selection in this context.

email: prabhanik@gmail.com

A STATISTICAL FRAMEWORK FOR INTEGRATIVE CLUSTERING ANALYSIS OF MULTI-TYPE GENOMIC DATA

Qianxing Mo*, Baylor College of Medicine
Ronglai Shen, Memorial Sloan-Kettering Cancer Center
Sijian Wang, University of Wisconsin, Madison
Venkatraman Seshan, Memorial Sloan-Kettering Cancer Center
Adam Olshen, University of California, San Francisco

We propose a framework for joint modeling of discrete and continuous data that arise from integrated genomic profiling experiments. The method integrates any combination of four different data types including binary (e.g., somatic mutation from exome sequening), categorical (e.g., DNA copy number gain, loss, normal), count (e.g., RNA-seq) and continuous (e.g., methylation and gene expression) data to identify the joint profiling patterns and genomic features that may be associated with tumor subtypes or phenotypes. The method provides a joint dimension reduction of complex genomic data to a few eigen features that are a mixture of linear and non-linear combinations of the original features and can be used for integrated visualization and cluster discovery. Genomic features are selected by applying a penalized likelihood approach with lasso penalty terms. We will use the Cancer Genome Atlas data and the cancer cell line encyclopedia data to illustrate its application.

email: qmo@bcm.edu

HIGH DIMENSIONAL SDEs COUPLED WITH MIXED-EFFECTS MODELING TECHNIQUES FOR DYNAMIC GENE REGULATORY NETWORK IDENTIFICATION

Iris Chen*, University of Rochester
Xing Qiu, University of Rochester
Hulin Wu, University of Rochester

Countless studies have been conducted to discover and investigate the complex system of gene regulatory networks (GRNs). Gene expression profiles from time-course experiments along with appropriate mathematical models and computational techniques will allow us to identify the dynamic regulatory networks. Stochastic differential equations (SDE) models can continuously capture the intrinsic dynamic gene regulation most comprehensively and meanwhile distinguish the stochastic system noise from measurement noise, which ordinary differential equations (ODE) models cannot accomplish. The noisy data and high dimensionality are the major challenges to all methodology development for GRN construction purposes. In addition, direct parameters estimation for high dimensional SDE is unattainable. We then proposed to utilize nonparametric mixed-effects clustering and smoothing and smoothly clipped absolute deviation (SCAD)-based variable selection techniques to efficiently obtain SDEs solution approximation and parameter estimation.

email: sinuiris@gmail.com

MODELING AND CHARACTERIZATION OF DIFFERENTIAL PATTERNS OF GENE EXPRESSION UNDERLYING PHENOTYPIC PLASTICITY USING RNA-Seq

Ningtao Wang*, The Pennsylvania State University
Yaqun Wang, The Pennsylvania State University
Zhong Wang, The Pennsylvania State University
Zuoheng Wang, Yale University
Kathryn J. Huber, The Pennsylvania State University
Jin-Ming Yang, The Pennsylvania State University
Rongling Wu, The Pennsylvania State University

RNA-Seq has increasingly played a pivotal role in measuring gene expression at an unprecedented precision and throughput and studying biological functions by linking differential patterns of expression with signal changes. A typical RNA-Seq experiment is to count transcript reads of all genes at two or more treatments and compare the difference of gene expression between the treatments. However, a challenge remains in cataloguing gene expression into distinct groups and interpreting differential patterns with biological functions. Here we address this challenge by developing a mechanistic model for gene clustering based on distinct patterns of gene expression in response to environmental change. The model was founded on a mixture likelihood of Poisson-distributed transcript read data, with each mixture component specified by the Skellam function. By estimating and comparing the amount of gene expression in each environment, the model allows the test of how genes alter their expression in response to environment and how different genes interact with each other in the responsive process. The statistical properties of the model were investigated through computer simulation. The new model provides a powerful tool for studying the plastic pattern of gene expression across different environments measured by RNA-Seq.

email: nxw5034@psu.edu

23. BIOSTATISTICAL METHODS IN FORENSICS, LAW AND POLICY

STATISTICAL METHODS FOR SIGNAL DETECTION IN LONGITUDINAL OBSERVATIONAL DRUG SAFETY DATA

Ram C. Tiwari*, U.S. Food and Drug Administration
Lan Huang, U.S. Food and Drug Administration
Jyoti N. Zalkikar, U.S. Food and Drug Administration

There are several statistical methods available for the signal detection in the FDA's Adverse Events Reporting System (AERS) database. These methods include the Proportional Odds Ratio (PRR), Multi-Gamma Poisson Shrinker (MGPS) and Bayesian Confidence Propagation Neural Network (BPCNN), among others. The AERS database consists of spontaneous adverse reports submitted directly from individuals, hospitals, and healthcare providers on drugs that are in the market after their approval, and as such the database may contain multiple reports on the same drug-adverse event combination from multiple sources. Therefore, the database does not have the information on the true (denominator) population size. As a result, the signals detected from these methods are exploratory or hypothesis-generating, and the data mining approach is called passive surveillance. Recently, Huang et al. (Jour. Amer. Stat. Assoc., 2011, 1230-1241) developed a likelihood ratio test for signal detection that assumes that the number of reports for a drug- adverse-event combination follows Poisson distribution. We present the LRT method and its extension for longitudinal observational and clinical exposure-based safety data. The performance of these tests using simulated data is evaluated and the methods are applied to the AERS data and to an exposure-based clinical safety data.

email: ram.tiwari@fda.hhs.gov

THE MATRIXX INITIATIVES V. SIRACUSANO CASE AND THE STATISTICAL ANALYSIS OF ADVERSE EVENT DATA

Joseph L. Gastwirth*, George Washington University

In the Matrixx Initiatives v Siracusano case, the Supreme Court ruled that information about the number of adverse events occurring to users of a drug need not be sufficiently numerous to reach statistically significance before they should be disclosed to potential investors. Surprisingly, none of the opinions or briefs presented a statistical analysis of the data. This talk describes a statistical method for comparing the proportion of individuals with a particular adverse event who had used the product under investigation with the market share of the product and concludes that there was a statistically significant excess of cases of loss of smell in users of Zicam nasal spray made by Matrixx. Because adverse event reports are not based on a random sample, a sensitivity analysis was carried, which showed that even if users were twice as likely to take the medicine as users of similar medicines and twice as likely to report a problem statistical significance remains. Monitoring adverse events is extremely important for protecting the public as it the studies of the

effectiveness of the medicine were too small to detect a small but important risk to users. If time permits, other aspects of the case that support the Court's decision will be noted.

email: jlgast@gwu.edu

STATISTICAL EVALUATION OF THE WEIGHT OF FINGERPRINT EVIDENCE: LEGAL PERSPECTIVE OF THE BENEFITS AND LIMITATIONS OF FINGERPRINT STATISTICAL MODELS

Cedric Neumann*, The Pennsylvania State University and Two's Forensics LLC

The first statistical model for the quantification of fingerprint evidence was proposed by Sir Francis Galton as early as 1892. Since then, a variety of models have been developed and proposed by statisticians, mathematicians and forensic scientists. Design shortcomings and operational constraints have prevented these models from being used by the forensic community to weight and report fingerprint evidence. Thus, for more than 100 years, fingerprint examiners have used heuristic processes to evaluate their evidence and form conclusions on the identity of the source of latent prints recovered from crime scenes. Recently, the use of more transparent models for the inference of the source of DNA evidence has resulted in numerous challenges of the scientific foundations of fingerprint examination. In turn, these challenges have led to a renewed interest in the development of statistical models for the evaluation of fingerprint evidence. Similarly to DNA evidence 25 years ago, fingerprint statistical models need to transition from the scientific to the legal arena. This paper will present a model developed at the Pennsylvania State University and validated using a 3M Cogent AFIS supported by more than 7,000,000 fingerprints, and will consider the elements that need to be in place to ensure a successful deployment of these models in forensic practice.

email: czn2@psu.edu

CASE COMMENTS ON ADAMS V. PERRIGO: ADOPTING A WEAKER CRITERION FOR BIOEQUIVALENCE IN PATENT INFRINGEMENT CASES THAN THE ONE IN APPROVING NEW DRUGS BY FDA

Qing Pan*, George Washington University

The concept of bioequivalence of two drugs in infringement cases may differ from the requirements used for drug approval by the FDA. The court with recent infringement case Adams v. Perrigo examined three different definitions of bioequivalence, which were presented by different parties. The statistical properties of those definitions are explored and evaluated. Our results support the appellate court's decision of less stringent requirements for bioequivalence in infringement cases.

email: qpan@gwu.edu

24. BRIDGING TO STATISTICS OUTSIDE THE PHARMACEUTICAL INDUSTRY: CAN WE BE MORE EFFICIENT IN DESIGNING AND SUPPORTING CLINICAL TRIALS?

CROSS-FERTILIZATION OF STATISTICAL DESIGNS AND METHODS BETWEEN BIOPHARMA AND OTHER INDUSTRIES

Jose C. Pinheiro*, Janssen Research & Development
Chyi-Hung Hsu, Janssen Research & Development

Because of its highly regulated nature, the biopharmaceutical industry has been, by and large, fairly insular when it comes to statistical designs and methods. Statistical approaches developed in other fields, such as engineering and manufacturing, rarely make their way into mainstream drug development applications. Likewise, methods developed within the biopharmaceutical context, such as group sequential designs, often do not find much application beyond their original target area. This lack of cross-industry synergism on the methodological front results in a considerable loss of efficiency in statistical knowledge sharing. This talk will discuss and illustrate opportunities for cross-industry application of statistical designs and methods, from other industries to biopharm and the other way around.

email: jpinhei1@its.jnj.com

CLINICAL TRIALS: PREDICTIVE ENROLLMENT MODELING AND MONITORING

Valerii Fedorov*, Quintiles

In recent years there is a significant trend to extend the arsenal of mathematical/statistical methods for design and logistic of clinical. Amazingly that most of these methods can be traced to the results that were known in communication engineering, reliability theory, quality control, etc. long ago but just very recently penetrated the disciplinary barriers to be used in pharmaceutical industry. For instance, the model based on mixtures of distributions just recently appeared in studies on predictive enrollment very much repeating what was done almost a century ago in communication theory. Similar examples can be found in risk based monitoring of multicenter clinical trials that are calling for the weak signal detection techniques or the use of sampling plans well established in manufacturing for decades. I will survey a few publications related to the above problem and discuss some potential extensions. Results for predictive modeling will be based on the facts well known in the Bayesian world and which are borrowed from what was developed in other research areas.

email: valerii.fedorov@quintiles.com

MULTI-ARM ADAPTIVE DESIGNS FOR PHASE II TRIALS IN RECURRENT GLIOBLASTOMA

Lorenzo Trippa*, Dana-Farber Cancer Institute

In recent years there has been a relevant increase in the number of clinical trials and putative antiangiogenic treatments for recurrent gliomas. A substantial part of these are single arm trials. I will consider response adaptive designs comparing a control with several novel treatments. These studies are designed to progressively increase the randomization probabilities for treatments which, on the basis of the data generated in the trial, show evidence of efficacy. We compare Bayesian adaptive randomization with alternative designs including two-arm and multi-arm balanced designs. The randomization probabilities are modified adaptively by sequentially updating a Bayesian model for the outcome distributions in each arm. The probability that a patient is assigned to a specific arm depends on the updated probability (at enrollment) that the corresponding treatment is effective. Our comparison focuses on realistic scenarios defined by using historical data, including progression free survival and overall survival, from recent trials. This study quantifies advantages and disadvantages of multi-arm Bayesian adaptive trials by means of a systematic assessment of the operating characteristics and suggests conclusions that can guide design on currently planned trials in glioma.

email: ltrippa@jimmy.harvard.edu

25. NEW ADVANCES IN FUNCTIONAL DATA ANALYSIS WITH APPLICATION TO MENTAL HEALTH RESEARCH

FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS FOR MULTIVARIATE FUNCTIONAL DATA

Chongzhi Di*, Fred Hutchinson Cancer Research Center

Functional data is becoming increasingly common in medical studies and health research. As a key technique for such data, functional principal component analysis (FPCA) was designed for a sample of independent functions. In this talk, we extend the scope of FPCA to multivariate functional data. We present several approaches to exploit the hierarchical structure of covariance operators at between and within subject levels, and extract dominating modes of variations at each level.

email: cdi@fhcrc.org

SPLINE CONFIDENCE ENVELOPES FOR COVARIANCE FUNCTION IN DENSE FUNCTIONAL/LONGITUDINAL DATA

Guanqun Cao*, Auburn University
Li Wang, University of Georgia
Yehua Li, Iowa State University
Lijian Yang, Soochow University and Michigan State University

We consider nonparametric estimation of the covariance function for dense functional data using tensor product B-splines. The proposed estimator is computationally more efficient than the kernel-based ones. We develop both local and global asymptotic distributions for the proposed estimator, and show that our estimator is as efficient as an 'oracle' estimator where the true mean function is known. Simultaneous confidence envelopes are developed based on asymptotic theory to quantify the variability in the covariance estimator and to make global inferences on the true covariance. Monte Carlo simulation experiments provide strong evidence that corroborates the asymptotic theory. Two real data examples on the near infrared spectroscopy data and speech recognition data are also provided to illustrate the proposed method.

email: gzc0009@auburn.edu

POINTWISE DEGREES OF FREEDOM AND MAPPING OF NEURODEVELOPMENTAL TRAJECTORIES

Philip T. Reiss*, New York University and Nathan Kline Institute
Lei Huang, Johns Hopkins University
Huaihou Chen, New York University

A major goal of biological psychiatry is to use brain imaging data to map trajectories of normal and abnormal development. In many applications the data may be viewed as functional responses that we would like to relate to age. More concretely, we are given a set of curves derived from individuals of different ages, where each person's curve represents a quantity of neuroscientific interest measured along a set of brain locations. The objective is to fit the quantity of interest as a smooth function of age at each location, while appropriately borrowing strength across locations. This talk will present a notion of pointwise degrees of freedom that provides a unifying framework for some existing approaches, and also motivates several new ones. The methods will be illustrated by application to a study of white matter microstructure in the corpus callosum.

email: phil.reiss@nyumc.org

VARIABLE SELECTION IN FUNCTIONAL LINEAR MODELS

Yihong Zhao*, New York University Medical Center

Variable selection plays an important role in high dimensional statistical modeling. We adopt the ideas of different screening strategies in high dimension linear models to linear models with functional predictors. We propose two classes of screening approaches: screening by variable importance and screening by stability. Finite sample performance of the proposed screening procedures is assessed by Monte Carlo simulation studies.

email: yz2135@caa.columbia.edu

26. SELECTION IN HIGH-DIMENSIONAL ANALYSIS

CLASSIFICATION RULE OF FEATURE AUGMENTATION AND NONPARAMETRIC SELECTION IN HIGH DIMENSIONAL SPACE

Jianqing Fan*, Princeton University
Yang Feng, Columbia University
Xin Tong, Massachusetts Institute of Technology

In this paper, we propose a new classification rule through feature augmentation and nonparametric selection (FANS) for high dimensional problems, where the number of features is comparable or much larger than the sample size. FANS follows a two step procedure. In the first step, marginal class conditional densities are estimated nonparametrically. In the second step, we invoke penalized logistic regression taking as input features the estimated log ratios of the marginal class conditional densities. A variant of FANS takes original features together with transformed features when applying penalized logistic regression. In theoretical derivations, it is assumed that the log ratios of class conditional densities can be written as a linear combination of log ratios of marginal class conditional densities. An oracle inequality regarding the risk is developed for FANS. The FANS model has rich interpretations. For example, it includes two class Gaussian models as a special case, and it can be thought of as an extension to nonparametric Naive Bayes. Also it is closely related to generalized additive models. In numerical analysis, we will compare FANS to these competing methods, so as to provide some guidelines about the best application domains of each method. Real data analysis demonstrates that FANS performs very competitively in bench mark email spam data and gene expression data. Also, the algorithm for FANS is extremely fast through parallel computing.

email: jqfan@princeton.edu

HIGH-DIMENSIONAL SPARSE ADDITIVE HAZARDS REGRESSION

Runze Li, The Pennsylvania State University
Jinchi Lv, University of Southern California

High-dimensional sparse modeling with censored survival data is of great practical importance, as exemplified by modern applications in high-throughput genomic data analysis and credit risk analysis. In this article, we propose a class of regularization methods for simultaneous variable selection and estimation in the additive hazards model, by combining the nonconcave penalized likelihood approach and the pseudoscore method. In a high-dimensional setting where the dimensionality can grow fast, polynomially or nonpolynomially, with the sample size, we establish the weak oracle property and oracle property under mild, interpretable conditions, thus providing strong performance guarantees for the proposed methodology. Moreover, we show that the regularity conditions required by the L1 method are substantially relaxed by a certain class of sparsity-inducing concave penalties. As a result, concave penalties such as the smoothly clipped absolute deviation (SCAD), minimax concave penalty (MCP), and smooth integration of counting and absolute deviation (SICA) can significantly improve on the L1 method and yield sparser models with better prediction performance. We present a coordinate descent algorithm for efficient implementation and rigorously investigate its convergence properties. The practical utility and effectiveness of the proposed methods are demonstrated by simulation studies and a real data example. This is a joint work with Wei Lin.

email: jinchilv@marshall.usc.edu

ULTRAHIGH DIMENSIONAL TIME COURSE FEATURE SELECTION

Peirong Xu, China East Normal University
Lixing Zhu, Hong Kong Baptist University
Yi Li*, University of Michigan

Statistical challenges arise from modern biomedical studies that produce time course genomic data with ultrahigh dimensions. For example, in a renal cancer study, the pharmacokinetic measures of a tumor suppressor (CCI-779) and expression levels of 12,625 genes were measured for each of 39 patients on 3 scheduled time points, with the goal of identifying predictive genes for pharmacokinetics over the time course. The resulting dataset defies analysis even with regularized regression. Although some remedies have been proposed for both linear and generalized linear models, there are virtually no solutions in the time course setting. As such, we propose a GEE-based screening procedure that only pertains to the specifications of the first two marginal moments and a working correlation structure. The new procedure is robust with respect to the mis-specification of correlation. It effectively reduces dimensionality of covariates and merely requires a single evaluation of GEE functions instead of fitting separate marginal models, as

often adopted by the existing methods. Our procedure enjoys computational and theoretical readiness, which is further verified via intensive Monte Carlo simulations. We also apply the procedure to analyze the aforementioned renal cancer study.

email: yili@umich.edu

AUTOMATIC STRUCTURE RECOVERY FOR ADDITIVE MODELS

Yichao Wu*, North Carolina State University
Len Stefanski, North Carolina State University

We propose an automatic structure recovery scheme for additive models. The structure recovery is based on a backfitting algorithm coupled with local polynomial smoothing, in conjunction with a new kernel-based variable selection strategy. The automatic structure recovery method produces estimates of the set of noise predictors, the sets of predictors that contribute polynomially at different orders up to any given order, and the set of predictors that contribute beyond polynomially. Asymptotic consistency of the method is proved. An extension to partially linear models is also described. Finite-sample performance of the proposed methods are illustrated via Monte Carlo studies and real data examples.

email: wu@stat.ncsu.edu

27. STATISTICS OF ENVIRONMENTAL HEALTH: CONSIDERING SPATIAL EFFECTS AND VARIOUS SOURCES OF POLLUTANT EXPOSURE ON HUMAN HEALTH OUTCOMES

BAYESIAN MODELS FOR CUMULATIVE SPATIAL-TEMPORAL RISK ASSESSMENT

Catherine Calder*, The Ohio State University
David Wheeler, Virginia Commonwealth University

In environmental epidemiology studies, environmental exposure is often determined by the spatial location of residence at the time of a health outcome or diagnosis (e.g., levels of pollution around an individual's home at the time of cancer diagnosis). However, due to the residential mobility of individuals and the potential long lag time between exposure to a relevant risk factor and an outcome such as diagnosis of cancer, it is reasonable to expect that historical residential locations of individuals may be more relevant in certain environmental health studies. To account for the spatio-temporal variation in environmental exposure due to residential mobility within a case-control framework, we develop Bayesian spatio-temporal distributed lag models to explore cumulative spatial-temporal risk to environmental toxicants based on individual residential histories. Our modeling framework allows us to simultaneously quantify the effects of environmental exposure on disease risk and understand disease latency periods. We illustrate our framework using data from a case-study of non-Hodgkin lymphoma (NHL) incidence within Los Angeles County, one of four centers in a population-based case-control study.

email: calder@stat.osu.edu

ON THE USE OF A PM_2.5 EXPOSURE SIMULATOR TO EXPLAIN BIRTHWEIGHT

Veronica J. Berrocal*, University of Michigan
Alan E. Gelfand, Duke University
David M. Holland, U.S. Environmental Protection Agency
Marie Lynn Miranda, University of Michigan

In relating pollution to birth outcomes, maternal exposure has usually been described using monitoring data. Such characterization provides a misrepresentation of exposure as (i) it does not take into account the spatial misalignment between an individual's residence and monitoring sites, and (ii) it ignores the fact that individuals spend most of their time indoors and typically in more than one location. In this paper, we break with previous studies by using a stochastic simulator to describe personal exposure (to particulate matter) and then relate simulated exposure at the individual level to the health outcome (birthweight) rather than aggregating to a selected spatial unit. We propose a hierarchical model that, at the first stage, specifies a linear relationship between birthweight and personal exposure, adjusting for individual risk factors and introduces random spatial effects for the census tract of maternal residence. At the second stage, we specify the distribution of each individual's personal exposure using the empirical distribution yielded by the stochastic simulator as well as a model for the spatial random effects.

email: berrocal@umich.edu

TIME SERIES ANALYSIS OF AIR POLLUTION AND HEALTH ACCOUNTING FOR SPATIAL EXPOSURE UNCERTAINTY

Howard Chang*, Emory University
Yang Liu, Emory University
Stefanie Sarnat, Emory University
Brian Reich, North Carolina State University

Exposure assessment in air pollution and health studies routinely utilize measurements from outdoor monitoring network which has limited spatial coverage. Moreover, ambient concentration may not reflect human exposure to outdoor pollution since individuals spend the majority of their time indoors. Exposure uncertainty can arise from spatial variation in air pollution concentration, as well as spatial variations in population or environmental characteristics that contribute to differential exposure. We will describe a time series study of fine particulate matter and emergency department visits in Atlanta. This analysis accounts for three well-recognized sources of exposure measurement error attributed to: (1) spatial variation in ambient concentration, (2) spatial variation in exposure-concentration relationship, and (3) spatial aggregation of health outcome. This is accomplished by incorporating additional data sources to supplement monitor measurements in a unified statistical modeling framework. Specifically, we first utilize remotely sensed data and process-based model output to obtain spatially-resolved concentration predictions. These predictions are then combined with data from stochastic exposure simulators to obtain estimated personal exposures.

email: howard.chang@emory.edu

CLIMATE CHANGE AND HUMAN MORTALITY

Richard L. Smith*, University of North Carolina, Chapel Hill and SAMSI
Ben Armstrong, London School of Hygiene and Tropical Medicine
Tamara Greasby, National Center for Atmospheric Research
Paul Kushner, University of Toronto
Joel Schwartz, Harvard School of Public Health
Claudia Tebaldi, Climate Central

The possible impacts of climate change include human health through a variety of mechanisms, including the direct effect of increasing temperatures on mortality and morbidity, the indirect effect through increased air pollution, and others including diarrhea, floods, malaria and malnutrition. In this talk we concentrate on the direct effects on mortality, which may be estimated through methods already well developed in epidemiology in the context of air pollution. In particular, we show some preliminary analyses based on the National Morbidity and Mortality Air Pollution Study (NMMAPS) dataset. We then discuss possible forward projections based on climate models, such as those documented by the North American Regional Climate Change Assessment Program (NARCCAP). We discuss a number of other issues, such as how to combine different sources of climate data, how to deal worth the effect of flu cycles and other seasonal influences, and the likelihood of adaptation to increasing temperatures.

email: rls@email.unc.edu

28. STATISTICAL ANALYSIS OF DYNAMIC MODELS: THEORY AND APPLICATION

MULTISTABILITY AND STATISTICAL INFERENCE OF DYNAMICAL SYSTEM

Wing H. Wong*, Stanford University
Arwen Meister, Stanford University
Henry Y. Li, Stanford University
Bokyung Choi, Stanford University
Chao Du, Stanford University

We discuss the use of dynamical systems to model gene regulatory processes in the cell. To be useful the models must be nonlinear. Our approach is to use gene perturbation to drive cells into different equilibrium states and then perform measurements on these states. This approach may be used to reconstruct a nonlinear system under minimal assumptions on its functional form. Finally, challenges related to the handling of intrinsic noise will be briefly discussed.

email: whwong@stanford.edu

DYNAMIC NETWORK MODELLING: LATENT THRESHOLD APPROACH

Mike West*, Duke University
Jouchi Nakajima, Duke University and Bank of Japan

The recently introduced concept of dynamic latent thresholding has been demonstrably valuable in a range of studies applying dynamic models in time series analysis and forecasting. Several recent applications include studies in finance and econometrics where the approach induces improved forecasts, resulting decisions and model interpretations. The ideas and models incorporating dynamic latent threshold mechanisms are broadly relevant in areas beyond these, including the biomedical sciences. We explore some of the potential here in studies of multivariate EEG time series in experimental neuropsychiatry, where time-varying vector autoregressions endowed with dynamic, probabilistic latent threshold mechanisms lead to improved model fits and interpretations relative to standard models, and lead immediately into a novel framework for statistical modelling of dynamic networks. We discuss the models and ideas of dynamic networks involving time-varying, feed-forward/back interconnections between network nodes. The latter allows for contemporaneous and/or lagged links between nodes to appear/disappear over time, as well as for formal estimation of time-varying weightings/relevances of links when present.

email: mw@stat.duke.edu

DATA-DRIVEN AUTOMATIC DIFFERENTIAL EQUATION CONSTRUCTOR WITH APPLICATIONS TO DYNAMIC BIOLOGICAL NETWORKS

Hulin Wu*, University of Rochester School of Medicine and Dentistry

Many systems in engineering and physics can be represented by differential equations, which can be derived from well-established physics laws and theories. However, currently no laws or theories exist to deduce exact quantitative relationships/interactions in biological world. It is unclear whether the biological systems follow a mathematical representation such as the differential equations, similar to that for a physics system. Fortunately, recent advances in cutting-edge biomedical technologies allow us to generate intensive high-throughput data to gain insights into biological systems. It is badly needed to develop statistical methods to test whether a biological system follows a mathematical representation based on experimental data. In this talk, I will present and discuss how to construct data-driven differential equations (ODE) to describe biological systems, in particular for dynamic gene regulatory network systems. The ODE models allow us to quantify both positive and negative regulations as well as feedback effects. We propose to combine the high-dimensional variable selection approaches and ODE model estimation methods to construct the ODE models based on experimental data. Application examples from biomedical studies will be presented to illustrate the proposed methodologies.

email: Hulin_Wu@urmc.rochester.edu

FAST ANALYSIS OF DYNAMIC SYSTEMS VIA GAUSSIAN EMULATOR

Samuel Kou*, Harvard University

Dynamic systems are used in modeling diverse behaviors in a wide variety of scientific areas. Current methods for estimating parameters in dynamic systems from noisy data are computationally intensive (for example, relying heavily on the numerical solutions of underlying differential equations). We propose a new inference method by creating a system driven by a Gaussian process to mirror the dynamic system. Auxiliary variables are introduced to connect this Gaussian system to the real dynamic system; and a sampling scheme is introduced to minimize the 'distance' between these two systems iteratively. The new inference method also covers the partially observed case in which only some components of the dynamic system are observed. The method offers a drastic saving of computational time and fast convergence while still retaining high estimation accuracy. We will illustrate the method by numerical examples.

email: kou@stat.harvard.edu

29. COMPLEX SURVEY METHODOLOGY AND APPLICATION

TWO-STAGE BENCHMARKING IN SMALL AREA ESTIMATION

Malay Ghosh*, University of Florida
Rebecca Steorts, University of Florida

The paper considers two-stage benchmarking in the context of small area estimation. A decision theoretic approach is used with asingle weighted squared error loss function that combines the loss at the domain-level and the area-level without any specific distributional assumptions. We consider this loss function while benchmarking the weighted means at each level or both the weighted means and weighted variability at the domain-level (under special conditions). We also provide multivariate versions of these results. Finally, we analyze the behavior of our methods using a study from the National Health Interview Survey (NHIS) from 2000, which estimates the proportion of people that do not have health insurance for many domains of the Asian subpopulation. We also perform a simulation study to look at the performance of the averaged squared errors of the various estimators of interest.

email: ghoshm@stat.ufl.edu

INFERENCE FOR FINITE POPULATION QUANTILES OF NON-NORMAL SURVEY DATA USING BAYESIAN MIXTURE OF SPLINES

Qixuan Chen*, Columbia University
Xuezhou Mao, Columbia University
Michael R. Elliott, University of Michigan
Roderick JA Little, University of Michigan

Non-normally distributed data are common in sample surveys. We propose a robust Bayesian model-based estimator for finite population quantiles of non-normal survey data in probability-proportional-to-size sampling. We assume that the probability of inclusion is known for all the units in the finite population. The non-normal distribution of the continuous survey variable is approximated using a mixture of normal distributions, in which both the mean and the variance of the survey variable are modeled as spline functions of the inclusion probabilities in each mixture component. A full Bayesian approach using the Markov chain Monte Carlo method is developed to obtain the posterior distribution of the finite population quantiles. We compare our proposed estimator with alternative estimators using simulations based on artificial data as well as a real finite population.

email: qc2138@columbia.edu

WHAT SURVEY AND MAINSTREAM STATISTICIANS ARE LEARNING FROM EACH OTHER

Phillip S. Kott*, RTI International

Until relatively recently, survey statistics has mostly been concerned with the estimation of means, totals, and ratios among the members of a finite population. In mainstream statistics, by contrast, the goal has often been to test whether some hypothesized model of unit behavior could be justified by the available sample data. A less appreciated distinction between the two is that the latter has traditionally been concerned with drawing efficient inference from small samples, while the former was more concerned with robust inference from large samples. As a consequence, survey statisticians have developed robust techniques requiring few, if any, model assumptions. The need to draw valid inferences from complex survey data has brought survey and mainstream statisticians into contact with each other. Mainstream statisticians have been using tools borrowed from survey statistics to handle these issues. Similarly, survey statisticians are increasingly finding models useful in their treatment of survey nonresponse. In addition, many are questioning their reliance on asymptotic normality. Although most survey samples are large, they are not always large in every domain of interest. We will review some of these developments and peer a bit into the future in this talk.

email: pkott@rti.org

EVALUATIONS OF MODEL-BASED METHODS IN ANALYZING COMPLEX SURVEY DATA: A SIMULATION STUDY USING MULTISTAGE COMPLEX SAMPLING ON A FINITE POPULATION

Rong Wei*, National Center for Health Statistics, Centers for Disease Control and Prevention
Van L. Parsons, National Center for Health Statistics, Centers for Disease Control and Prevention
Jennifer D. Parker, National Center for Health Statistics, Centers for Disease Control and Prevention

Traditional design-based methods for complex-survey data often provide unreliable analyses when sample sizes are not sufficiently large to achieve approximate asymptotic sampling distributions for the direct estimators under consideration. Model-based estimation methods are often suggested as alternatives to compensate for data deficiencies. For this study, we consider some fundamental multi-level models whose features are used to account for survey weights and clustering. We examine their sampling distributions by means of a simulation of a complex-survey sample from a pseudo population. This pseudo population was developed from 9 years of National Health Interview Survey (NHIS) data, and it captures many features of geographical and household clustering within the true population. Multi-stage probability sampling and estimation methods consistent with those used in the NHIS are a major part of the simulation. Using SAS® procedures PROC MIXED and PROC GLIMMIX for the modeling, and SAS SURVEY procedures for the design-based approaches, the sampling and inferential properties of the two types of methods are compared.

email: rrw5@cdc.gov

30. DOSE-RESPONSE AND NONLINEAR MODELS

HIERARCHICAL DOSE-RESPONSE MODELING FOR HIGH-THROUGHPUT TOXICITY SCREENING OF ENVIRONMENTAL CHEMICALS

Ander Wilson*, North Carolina State University
David Reif, North Carolina State University
Brian Reich, North Carolina State University

High-throughput screening (HTS) of environmental chemicals is used to identify chemicals with high potential for adverse human health and environmental effects. Predicting physiologically-relevant activity with HTS data requires estimating the response of hundreds of chemicals across a battery of screening assays based on sparse dose-response data for each chemical-assay combination. Many standard dose-response methods are inadequate because they treat each curve as independent and under-perform when there are as few as seven observed responses. We propose two hierarchical Bayesian models, one parametric and one semiparametric, that borrow strength across chemicals and assays. Our methods directly parametrize the efficacy and potency of the responses as well as the probability of response. We use the ToxCast data from the U.S. EPA as motivation. We demonstrate that our hierarchical methods outperform independent curve estimation in a simulation study and provide more accurate estimates of the probability of response, efficacy, and potency and that our semiparametric method is robust to assumptions about the shape of the response. We use our semiparametric method to rank chemicals in the to ToxCast data by potency and compare untested chemicals with well-studies reference chemical to predict which chemicals may have related biological effects at lower doses.

email: anderwilson@gmail.com

ESTIMATING BROOD-SPECIFIC REPRODUCTIVE INHIBITION POTENCY IN AQUATIC TOXICITY TESTING

Jing Zhang*, Miami University
A. John Bailer, Miami University
James T. Oris, Miami University

Chemicals from effluents may impact the mortality, growth, or reproduction of organisms in the aquatic systems. A common endpoint analyzed in the reproduction toxicology is the total young produced by organisms exposed to toxicants. In the present study, we propose using two Bayesian hierarchical models to analyze the brood-specific reproduction count responses in aquatic toxicology experiments jointly and to estimate the brood-specific concentration associated with a specified level of reproductive inhibition. A simulation studies showed that the proposed models outperformed an approach where brood-specific responses were modeled separately. In particular this method provided potency estimates with smaller bias/variation and better coverage probability. The application of the proposed models is illustrated with an experiment where the impact of Nitrofen was studied.

e-mail: zhangj8@muohio.edu

A DIVERSITY INDEX FOR MODEL SELECTION IN THE ESTIMATION OF BENCHMARK AND INFECTIOUS DOSES VIA FREQUENTIST MODEL AVERAGING

Steven B. Kim*, University of California Irvine
Ralph L. Kodell, University of Arkansas for Medical Sciences
Hojin Moon, California State University, Long Beach

In chemical and microbial risk assessments, risk assessors fit dose-response models to high-dose data and extrapolate downward to risk levels in the range of 1% to 10%. Although multiple dose-response models may be able to fit the data adequately in the experimental range, the estimated effective dose corresponding to an extremely small risk can be substantially different from model to model. In this respect, using model averaging is more appropriate than relying on any single dose-response model in the calculation of a point estimate and a lower confidence limit for an effective dose. In model averaging, accounting for both data uncertainty and model uncertainty is crucial, but proper variance estimation is not guaranteed simply by increasing the number of models in a model space. A plausible set of models in model averaging can be characterized by good fits to the data and diversity surrounding the truth. We propose a diversity index for model selection which balances between goodness-of-fit and divergence among a set of parsimonious models. Tuning parameters in the diversity index control the size of the model space for model averaging. The proposed method is illustrated with two experimental data sets.

e-mail: steven.b.kim@gmail.com

TESTING FOR CHANGE POINTS DUE TO A COVARIATE THRESHOLD IN REGRESSION QUANTILES

Liwen Zhang*, Fudan University
Huixia Judy Wang, North Carolina State University
Zhongyi Zhu, Fudan University

We develop a new procedure for testing change points due to a covariate threshold in regression quantiles. The proposed test is based on the cumsum of the subgradient

of the quantile objective function and requires fitting the model only under the null hypothesis. The critical values can be obtained by simulating the Gaussian process that characterizes the limiting distribution of the test statistic. The proposed method can be used to detect change points at a single quantile level or across quantiles, and can accommodate both homoscedastic and heteroscedastic errors. Simulation results suggest that the proposed methods have more power in finite samples and higher computational efficiency than the existing likelihood-ratio-based method. A real example is given to illustrate the performances of our proposed methods.

e-mail: lzhang34@ncsu.edu

SEMIPARAMETRIC BAYESIAN JOINT MODELING OF A BINARY AND CONTINUOUS OUTCOME WITH APPLICATIONS IN TOXICOLOGICAL RISK ASSESSMENT

Beom Seuk Hwang*, The Ohio State University
Michael L. Pennell, The Ohio State University

Many dose-response studies collect data on correlated outcomes. For example, in developmental toxicity studies, uterine weight and presence of malformed pups are measured on the same dam. Joint modeling can result in more efficient inferences than independent models for each outcome. Most methods for joint modeling assume standard parametric response distributions. However, in toxicity studies, it is possible that response distributions vary in location and shape with dose, which may not be easily captured by standard models. We propose a semiparametric Bayesian joint model for a binary and continuous response. In our model, a kernel stick-breaking process (KSBP) prior is assigned to the distribution of a random effect shared across outcomes, which allows flexible changes in shape with dose shared across outcomes. The model also includes outcome-specific fixed effects to allow different location effects. In simulation studies, we found that the proposed model provides accurate estimates of toxicological risk when the data does not satisfy assumptions of standard parametric models. We apply our method to data from a developmental toxicity study of ethylene glycol diethyl ether.

e-mail: hwang.176@osu.edu

A SIGMOID SHAPED REGRESSION MODEL WITH BOUNDED RESPONSES FOR BIOASSAYS

HaiYing Wang, University of Missouri
Nancy Flournoy*, University of Missouri

We introduce a new sigmoid shaped regression model in which the response variable is bounded by two unknown parameters. A special case is a bounded alternative to the four parameter logistic model which is widely used in bioassay, nutrition, genetics, calibration and agriculture. When responses are bounded but the bounds are unknown, our model better reflects the data-generating mechanism. Complications arise because the likelihood function is unbounded, and the global maximizers are not consistent estimators of unknown parameters. Although the two sample extremes, the smallest and the largest observations, are consistent estimators for the two unknown boundaries, they have slow convergence rate and are asymptotically biased. Bias corrected estimators are developed in the one sample case; but they do not obtain the optimal convergence rate. We recommend using the local maximizers of the likelihood function, i.e., the solution to the likelihood equations. We prove, with probability approaching one, there exists a solution to the likelihood equation that is consistent at the rate of the square root of the sample size and it is asymptotically normally distributed. Examples are provided and design issues discussed.

e-mail: flournoyn@missouri.edu

MODEL SELECTION AND BMD ESTIMATION WITH QUANTAL-RESPONSE DATA

Edsel A. Pena*, University of South Carolina
Wensong Wu, Florida International University
Walter W. Piegorsch, University of Arizona
Webster R. West, North Carolina State University
Lingling An, University of Arizona

This talk will describe several approaches for estimating the benchmark dose (BMD) in a risk assessment study with quantal dose-response data and when there are competing model classes for the dose-response function. Strategies involving a two-step approach, a model-averaging approach, a focused-inference approach, and a nonparametric approach based on a PAVA-based estimator of the dose-response function are described and compared. Attention is raised to the perils involved in data "double-dipping" and the need to adjust for the model-selection stage in the estimation procedure. Simulation results are presented comparing the performance of five model selectors and eight BMD estimators. An illustration using a real quantal-response data set from a carcinogenecity study is provided.

e-mail: pena@stat.sc.edu

31. METHODS AND APPLICATIONS IN COMPARATIVE EFFECTIVENESS RESEARCH

COST-EFFECTIVENESS INFERENCE WITH SKEWED DATA

Ionut Bebu*, Infectious Disease Clinical Research Program, Uniformed Services University of the Health Sciences
George Luta, Georgetown University
Thomas Mathew, University of Maryland, Baltimore County
Paul A. Kennedy, Infectious Disease Clinical Research Program, Uniformed Services University of the Health Sciences
Brian Agan, Infectious Disease Clinical Research Program, Uniformed Services University of the Health Sciences

Evaluating the treatment effect while taking into account the associated costs is an important goal of cost-effectiveness analyses. Several cost-effectiveness measures have been proposed to quantify these comparisons, including the incremental cost-effectiveness ratio (ICER) and the incremental net benefit (INB). Various approaches have been proposed for constructing confidence intervals for ICER and INB, including parametric methods (e.g. based on the Delta method or on Fieller's method), nonparametric methods (e.g. various bootstrap methods), as well as Bayesian methods. Skewed data are usually the norm in cost-effectiveness analyses, and accurate parametric confidence intervals in this context are lacking. We constructed confidence intervals for both ICER and INB using the concept of a generalized pivotal quantity, which can be derived for various combinations of normal, lognormal, and gamma costs and effectiveness, with and without covariates. The proposed methodology is straightforward in terms of computation and implementation, and the resulting confidence intervals compared favorably with existing methods in a simulation study. Our approach is illustrated using three randomized trials.

email: ibebu@idcrp.org

CONSIDERING BAYESIAN ADAPTIVE DESIGNS FOR COMPARATIVE EFFECTIVENESS RESEARCH: REDESIGN OF THE ALLHAT TRIAL

Kristine R. Broglio*, Berry Consultants, LLC
Jason T. Connor, Berry Consultants, LLC and University of Central Florida College of Medicine

The REsearch in ADAptive methods for Pragmatic Trials (RE-ADAPT) project is funded by a National Heart Lung and Blood Institute (NHLBI) grant. The aim of RE-ADAPT is to demonstrate the design and conduct of Bayesian adaptive trials in a comparative effectiveness setting. We use the Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) as an example. ALLHAT was a comparative effectiveness trial for the

prevention of fatal coronary heart disease and non-fatal MI. This trial enrolled 42,418 patients, cost 135 million dollars, and lasted 8 years, yet, by the time ALLHAT ended, practice patterns had changed significantly and the clinical questions were no longer as relevant. We explore whether a Bayesian adaptive trial design would have met the primary objective more efficiently. Using only information available to the original trial designers in the early 1990s, we prospectively define seven Bayesian adaptive alternative designs for ALLHAT. We consider early stopping, adaptive randomization, and arm dropping. We use simulation to determine the operating characteristics of each design and to choose an optimal alternative design. On average, many of the alternative designs produce shorter trial durations with a higher proportion of patients being randomized to the most effective therapies.

email: kristine@berryconsultants.com

TESTING BAYESIAN ADAPTIVE TRIAL STRATEGIES IN CER: RE-EXECUTION OF ALLHAT

Jason T. Connor*, Berry Consultants, LLC and University of Central Florida College of Medicine
Kristin R. Broglio, Berry Consultants, LLC

The REsearch in ADAptive methods for Pragmatic Trials (RE-ADAPT) project aims to demonstrate how Bayesian adaptive trials might improve the efficiency of large-scale comparative effectiveness research projects using ALLHAT as a case study. ALLHAT was designed to compare four anti-hypertensive drugs and enrolled 42,418 patients, cost 135 million dollars, and lasted 8 years. Yet, by ALLHAT's conclusion, practice patterns had changed significantly and its clinical questions were no longer as relevant. We explore whether a Bayesian adaptive trial design would have met the primary objective more efficiently. Using the seven designs prespecified in the grant's Aim 1, in Aim 2 we conduct those seven trials using the actual ALLHAT data. We compare the various adaptive design components used in each and how they effect the proportion of patients randomized to the best therapies, the timing of the trial and the trial's final sample size. We discuss the pros and cons of using Bayesian adaptive trials for large-scale comparative effectiveness research projects.

email: jason@berryconsultants.com

EFFECT MODIFICATION BY POST-TREATMENT VARIABLES IN MENTAL HEALTH RESEARCH

Alisa J. Stephens*, University of Pennsylvania
Marshall M. Joffe, University of Pennsylvania

In comparative effectiveness research, interest lies in determining how the effect of an intervention varies with patient characteristics. Standard methods consider how the effect of a treatment or exposure is modified by variables measured at the time of treatment. We introduce Retrospective Structural Nested Mean Models (RSNMMs) to evaluate effect modification of treatment by a post-treatment covariate. Such post-treatment effect modification may be of interest in determining when to modify a treatment based on a patient's initial response. Our models differ from regular SNMMs in that causal effects of treatments may be modeled in subgroups defined by covariates that are measured after treatment. We discuss an estimation procedure for the case of binary outcomes modeled through a logit link and apply our methods to a randomized trial in mental health research.

email: alisaste@mail.med.upenn.edu

SENSITIVITY ANALYSIS FOR INSTRUMENTAL VARIABLES REGRESSION OF THE COMPARATIVE EFFECTIVENESS OF REFORMULATED ANTIDEPRESSANTS

Jaeun Choi*, Harvard Medical School
Mary Beth Landrum, Harvard Medical School
A. James O'Malley, Harvard Medical School

We evaluate the sensitivity of the results of an instrumental variable (IV) regression analysis of an observational study comparing reformulated and original formulations of antidepressant medications on the likelihood a patient discontinues treatment. We first investigate sensitivity to the geographical unit at which the IV and location-control dummy variables are defined. The analysis is motivated by the trade-off between the efficiency of IV and the level of control of unmeasured confounding variables that vary by area. We also assess sensitivity of the results to violation of the exclusion restriction assumption required for IV analysis to yield unbiased results by quantifying the extent to which the results would be expected to change as a function of the magnitude of the violation. The advantages of different ways of specifying the magnitude of the violation in an applied problem are described, evaluated and discussed.

email: choi@hcp.med.harvard.edu

ASSESSING THE CAUSAL EFFECT OF TREATMENT IN THE PRESENCE OF PARTIAL COMPLIANCE

Xin Gao*, University of Michigan
Michael R. Elliott, University of Michigan

To make drug therapy as effective as possible, patients are often put on an escalating dosing schedule. But patients may choose to take a lower dose because of side effects. Therefore, even if the dose schedule is randomized, the dose level received is a post-randomization variable, and comparison between the treatment arm and the control arm may no longer have a causal interpretation. We use the potential outcomes framework to define pre-randomization "principal strata" from the distribution of dose tolerance under treatment arm, with the goal of estimating the causal effect of treatment within the subgroups of the population who are able to tolerate a given level of treatment dose. Adverse events are included in the model to help identify subjects' principal strata membership. Inference is obtained by treating the outcomes, doses, and adverse events under the unobserved randomization arm as missing data and using multiple imputation in a Bayesian framework. Results from simulation studies imply that the proposed causal model provides valid inferences of the causal effect of treatment within principal strata under certain reasonable assumptions. We apply the proposed model to a randomized clinical trial with escalating dosing schedule for the treatment of interstitial cystitis and estimate the causal effects of treatment within principal strata.

email: xingao@umich.edu

TOO MANY COVARIATES AND TOO FEW CASES? A COMPARATIVE STUDY

Qingxia Chen*, Vanderbilt University
Yuwei Zhu, Vanderbilt University
Marie R. Griffin, Vanderbilt University
Keipp H. Talbot, Vanderbilt University
Frank E. Harrell, Vanderbilt University

For the multivariable logistic regression model, it is recommended to include at most 10-15 parameters per valid sample size m in order to reliably estimate the regression coefficients, where m is the number of cases or controls, whichever is less. This condition is, however, hard to be met even in a well designed study when the number of confounders is overwhelmed, rare disease is studied, and/or subgroup analysis is of interest. Extensive simulations were conducted to evaluate various existing methods including various propensity score methods (adjustment, stratify, weighting, or additional heterogeneity adjustment), and penalized regression models including ridge regression, LASSO, SCAD, and elastic net. The methods were evaluated in the setups which mimic our motivating clinical data and the results showed that the penalized logistic regression model with ridge penalty and the logistic regression model with propensity score adjustment outperform the other methods in our setting. There are some factors that will affect the choice between the aforementioned two methods: (a) exposure rate; (b) number of valid sample size per parameter; (c) ratio between cases and controls. We applied the methods to estimate the vaccine effectiveness in the motivating vaccine study.

email: cindy.chen@vanderbilt.edu

32. BAYESIAN METHODS

BAYESIAN GENERALIZED LOW RANK REGRESSION MODELS FOR NEUROIMAGING PHENOTYPES AND GENETIC MARKERS

Zakaria Khondker*, University of North Carolina, Chapel Hill and PAREXEL International
Hongtu Zhu, University of North Carolina, Chapel Hill
Joseph Ibrahim, University of North Carolina, Chapel Hill

We propose a Bayesian generalized low rank regression model (GLRR) for the analysis of both high-dimensional responses and covariates. This development is motivated by performing genome-wide searches for associations between genetic variants and brain imaging phenotypes. GLRR integrates a low rank matrix to approximate the high-dimensional regression coefficient matrix of GLRR and a dynamic factor model to model the high-dimensional covariance matrix of brain imaging phenotypes. Local hypothesis testing is developed to identify significant covariates on high-dimensional responses. Posterior computation proceeds via an efficient Markov chain Monte Carlo algorithm. A simulation study is performed to evaluate the finite sample performance of GLRR and its comparison with several competing approaches. We apply GLRR to investigate the impact of 10,479 SNPs on chromosome 9 on the volumes of 93 regions of interest (ROI) obtained from Alzheimer's Disease Neuroimaging Initiative (ADNI).

email: khondker@email.unc.edu

BAYESIAN ANALYSIS OF CONTINUOUS CURVE FUNCTIONS

Wen Cheng*, University of South Carolina, Columbia
Ian Dryden, University of Nottingham, UK
Xianzheng Huang, University of South Carolina, Columbia

We consider Bayesian analysis of continuous curve functions in 1D, 2D and 3D space. A fundamental aspect of the analysis is that it is invariant under a simultaneous warping of all the curves, as well as translation, rotation and scale of each individual. We introduce Bayesian models based on the curve representation named Square Root Velocity Function (SRVF) introduced by Srivastava et al. (2011, IEEE PAMI). A Gaussian process model for SRVF of curves is proposed, and suitable prior models such as Dirichlet process are employed for modeling the warping function as a Cumulative Distribution Function (CDF). Simulation from posterior distribution is via Markov chain Monte Carlo methods, and credibility regions for mean curves, warping functions as well as nuisance parameters are obtained. Special treatment needs to be applied when target curves are closed. We will illustrate the methodology with applications in 1D proteomics data, 2D mouse vertebra outlines and 3D protein secondary structure data.

email: chengwen1985@gmail.com

A BAYESIAN MODEL FOR IDENTIFIABLE SUBJECTS

Edward J. Stanek III*, University of Massachusetts, Amherst
Julio M. Singer, University of Sao Paulo, Brazil

Practical problems often involve estimating individual latent values based on data from a sample. We discuss an application where latent LDL cholesterol levels of women from an HMO are of interest. We use a Bayesian model with an exchangeable prior distribution that includes subject labels, and trace how the prior distribution is updated via the data to produce the posterior distribution. The prior distribution is specified for a finite population of women in the HMO assumed to arise from a larger superpopulation. The novel aspect is accounting for the labels in the prior. We illustrate this via an example, and show how the exchangeable prior distribution can be constructed for a finite population that arose from a superpopulation. Using data that consists of the (label, response) pair for a set of women, we illustrate how conditioning on the set of labels, the sequence of labels in the sample space, and the actual response impacts the change from the prior to the posterior distributions. In particular, we show that conditioning on the actual response, alters the distribution of latent values for the women in the data, but not for the remaining women in the population.

email: stanek@schoolph.umass.edu

A LATENT VARIABLE POISSON MODEL FOR ASSESSING REGULARITY OF CIRCADIAN PATTERNS OVER TIME

Sungduk Kim*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Paul S. Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Actigraphs are often used to assess circadian patterns in activity across time. Although there is a statistical literature on modeling the circadian mean structure, little work has been done in understanding variations in these patterns over time. The NEXT generation health study collects longitudinal actigraphs over a seven day period, where activity counts are observed at 30 second epochs. Exploratory analysis suggest that some individuals have very regular circadian patterns in activity, while others show marked variation in these patterns. We develop a latent variable Poisson model that characterizes irregularity in the circadian pattern through latent variables for circadian, stochastic, and individual variation. A parameterization is proposed for modeling covariate dependence on the degree of regularity in the circadian patterns over time. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed model are considered and compared via the deviance information criterion. The proposed methodology is motivated by and applied to the NEXT generation health study conducted by the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the Nation Institutes of Health in the summer of 2010.

email: kims2@mail.nih.gov

OVERLAP IN TWO-COMPONENT MIXTURE MODELS: INFLUENCE ON INDIVIDUAL CLASSIFICATION

José Cortiñas Abrahantes*, European Food Safety Authority
Geert Molenberghs, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium

The problem of modeling the distributions of the continuous scale measured values from a diagnostic test (DT) for a specific disease is often encounter in epidemiological studies, in which disease status of the animal might depend on characteristics such as herd, age and/or other disease-specific risk factors. Techniques to decompose observed DT values into their underlying components are of interest, for which mixture models offer a viable pathway to deal with classification of individual samples, and at the same time account for other factors influencing the individual classification. Mixture models have been frequently used as a clustering technique, but the classification performance of individual observations has been rarely discussed. A case study in salmonella was the motivating force to study classification performance based on mixture model. Simulations using different measures to quantify overlap of the components and with this the degree of separation between the components were carried out. The results provide insight into the potential problems that could occur when using mixture models to assign individual observations to specific component in the population. The measures of overlap prove useful to identify the potential ranges for each of the classification performance measures, once the mixture model is fitted.

email: jose.cortinasabrahantes@efsa.europa.eu

EMPIRICAL AND SMOOTHED BAYES FACTOR TYPE INFERENCES BASED ON EMPIRICAL LIKELIHOODS FOR QUANTILES

Ge Tao*, State University of New York at Buffalo
Albert Vexler, State University of New York at Buffalo
Jihnhee Yu, State University of New York at Buffalo
Nicole A. Lazar, University of Georgia
Alan Hutson, State University of New York at Buffalo

The Bayes factor, a practical tool of applied statistics, has been dealt with extensively in the literature in the context of hypothesis testing. The Bayes factor based on parametric likelihoods can be considered both as a pure Bayesian approach as well as a standard technique for computing P-values for hypothesis testing when the functional forms of the data distributions are known. In this article, we employ empirical likelihood methodology in order to modify Bayes factor type procedures for the nonparametric setting. The proposed approach is applied towards developing testing methods involving quantiles, which are commonly used to characterize distributions. Comparing quantiles thus provides valuable information; however, very few tests for quantiles are available. We present and evaluate one- and two-sample distribution-free Bayes factor type methods for testing quantiles based on indicators and smooth kernel functions. Although the proposed procedures are nonparametric, their asymptotic behaviors are similar to those of the classical Bayes factor approach. An extensive Monte Carlo study and real data examples show that the developed tests have excellent operating characteristics for one-sample and two-sample data analysis.

email: getao@buffalo.edu

MODELING UNCERTAINTY IN BAYESIAN CONSTRAINT ANALYSIS

Zhen Chen*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Michelle Danaher, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Anindya Roy, University of Maryland Baltimore County

We propose a new Bayesian framework to estimating constrained parameters when the constraints are subject to uncertainty. Based on the principle that uncertainty about constraints is a manifestation of uncertainty about the boundary of the constraint set, we treat the boundary as a random quantity and assign prior distribution for it. This makes it possible that the constraints can be potentially violated when they are contradicted by data. Furthermore, the degree of uncertainty can be subjectively controlled in the boundary prior. We show that the proposed framework, when compared to an existing one, has a logical underpinning principle, applies to more general problems, and allows easy estimation procedures.

email: chenzhe@mail.nih.gov

33. VARIABLE SELECTION PROCEDURES

ADAPTIVE COMPOSITE M-ESTIMATION FOR PARTIALLY OVERLAPPING MODELS

Sunyoung Shin*, University of North Carolina, Chapel Hill
Jason P. Fine, University of North Carolina, Chapel Hill
Yufeng Liu, University of North Carolina, Chapel Hill

This paper proposes a simultaneously penalized M-estimation with composite convex loss function. Composite loss function is a weighted linear combination of convex loss functions which have their own coefficients vector. Simultaneous sparsity and grouping penalty terms regularize the composite loss function. Penalization on a single coefficient for every loss function encourages sparsity on every loss function. Penalization on pairwise coefficients difference across loss functions encourages grouping within coefficients corresponding to one predictor. We establish the oracle properties of M-estimation with composite loss function. We demonstrate that simultaneously penalized M-estimation with composite loss function enjoys the oracle property. Selection and grouping consistency are achieved. Estimation efficiency of non-zero coefficients is improved by grouping the coefficients across loss functions. The grouping contributes efficiency of M-estimation. Simulation studies and real data analysis illustrate that our method has advantage over other existing methods.

email: sunyoung@live.unc.edu

FREQUENTIST CONFIDENCE INTERVALS FOR THE SELECTED TREATMENT MEANS

Claudio Fuentes*, Oregon State University
George Casella, University of Florida

Consider an experiment in which p independent treatments or populations pi_i, with corresponding unknown means theta_i are available and suppose that for every population we can obtain a random sample. In this context, researchers are sometimes interested in selecting the populations that give the largest sample means as a result of the experiment, and to estimate the corresponding population means theta_i's. In this talk, we present a frequentist approach to the problem and discuss how to construct confidence intervals for the mean of the selected population, assuming the populations pi_i are independent and normally distributed with a common variance sigma^2. This approach, based on minimization of the coverage probability, produces asymmetric confidence intervals that maintain the nominal coverage probability, taking into account the selection procedure. Extensions of this approach to the problem following selection of k populations will be discussed.

email: fuentesc@stat.oregonstate.edu

BAYESIAN SEMIPARAMETRIC RANDOM EFFECT SELECTION IN GENERALIZED LINEAR MIXED MODELS

Yong Shan*, University of South Carolina
Xiaoyan Lin, University of South Carolina
Bo Cai, University of South Carolina

Random effects selection in Generalized linear mixed models (GLMM) is challenging, especially when the random effects are assumed nonparametrically distributed. In this paper, we develop a unified Bayesian approach for the random effect selection for the GLMMs. Specifically, we assume the random effects arise from a Dirichlet process (DP) prior with normal base measure. A special Cholesky-type decomposition is applied to the base covariance in the DP prior, and then zero-inflated mixture priors are assigned to the components of the decomposition to achieve the random effect selection. For non-Gaussian data, a Laplace approximation to the likelihood function is relied to apply the proposed MCMC algorithm. The performance of our proposed approach is investigated by using simulated data and real life data.

email: shany@email.sc.edu

RANDOM EFFECTS SELECTION IN BAYESIAN ACCELERATED FAILURE TIME MODEL WITH CORRELATED INTERVAL CENSORED DATA

Nusrat Harun*, University of South Carolina
Bo Cai, University of South Carolina

In many medical problems that collect multiple observations per subject, the time to an event is often of interest. Sometimes, the occurrence of the event can be recorded at regular intervals leading to interval censored data. It is further desirable to obtain the most parsimonious model in order to increase predictive power and to obtain ease of interpretation. Variable selection and often random effect selection in case of clustered data becomes crucial in such applications. We propose a Bayesian method for random effects selection in mixed effects accelerated failure time models. The proposed method relies on Cholesky decomposition on the random effects covariance matrix and the parameter expansion method for the selection of random effects. The Dirichlet prior is used to model the uncertainty in the random effects. The error distribution for the AFT model has been specified using a Gaussian mixture to allow flexible error density and prediction of the survival and hazard functions. We demonstrate the model using extensive simulations and the Signal Tandmobiel Study®.

email: nharun@mdanderson.org

BAYESIAN SEMIPARAMETRIC VARIABLE SELECTION WITH APPLICATION TO DENTAL DATA

Bo Cai*, University of South Carolina
Dipankar Bandyopadhyay, University of Minnesota

Normality assumption is typically adopted for random effects in repeated and longitudinal data analysis. However, such an assumption is not always realistic as random effects could follow any distribution. The violation of normality assumption may lead to potential biases of the estimates, especially when variable selection is taken into account. On the other hand, flexibility of nonparametric assumptions (e.g. Dirichlet process) may potentially cause centering problems which lead to difficulty of interpretation of effects and variable selection. Motivated by this problem, we propose a Bayesian method for fixed and random effects selection in nonparametric random effects models. We model the regression coefficients via centered latent variables which are distributed as probit stick-breaking (PSB) scale mixtures (Pati and Dunson, 2011). By using the mixture priors for centered latent variables along with covariance decomposition, we can avoid the aforementioned problems and allow fixed and random effects to be effectively selected in the model. We demonstrate advantages of the proposed approach over the other methods in the simulated example. The proposed method is further illustrated through an application to dental data.

email: bcai@sc.edu

STRUCTURED VARIABLE SELECTION WITH Q-VALUES

Tanya P. Garcia*, Texas A&M University
Samuel Mueller, University of Sydney
Raymond J. Carroll, Texas A&M University
Tamara N. Dunn, University of California, Davis
Anthony P. Thomas, University of California, Davis
Sean H. Adams, U.S. Department of Agriculture, Agricultural Research Service Western Human Nutrition Research Center
Suresh D. Pillai, Texas A&M University
Rosemary L. Walzem, Texas A&M University

When some of the regressors can act on both the response and other explanatory variables, the already challenging problem of selecting variables when the number of covariates exceeds the sample size becomes more difficult. A motivating example is a metabolic study in mice that has diet groups and gut microbial percentages that may affect changes in multiple phenotypes related to body weight regulation. The data have more variables than observations and diet is known to act directly on the phenotypes as well as on some or potentially all of the microbial percentages. Interest lies in determining which gut microflora influence the phenotypes while accounting for the direct relationship between diet and the other variables. A new methodology for variable selection in this context is presented that links the concept of q-values from multiple hypothesis testing to the recently developed weighted Lasso.

email: tpgarcia@stat.tamu.edu

VARIABLE SELECTION IN SEMIPARAMETRIC TRANSFORMATION MODELS FOR RIGHT CENSORED DATA

Xiaoxi Liu*, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill

There is limited work on variable selection for general transformation models with censored data. Existing methods use either estimating functions or ranks so they are computationally intensive and inefficient. In this work, we propose a computationally simple method for variable selection in general transformation models. The proposed algorithm reduces to maximizing a weighted partial likelihood function within an adaptive LASSO framework. It includes both proportional odds model and proportional hazard model as special cases and easily incorporate time-dependent covariates. We establish the asymptotic properties of the proposed method, including selection consistent and semiparametric efficiency of the post-selection estimator. Our simulations demonstrate a good small-sample performance of the proposed method and indicate that the variable selection result is robust even if transformation function is misspecified. A real data analysis shows that the proposed method outperforms an external risk score method in future prediction.

email: xiaoxi1@unc.edu

34. CLUSTERED DATA METHODS

VIRAL GENETIC LINKAGE ANALYSES IN THE PRESENCE OF MISSING DATA

Shelley H. Liu*, Harvard School of Public Health
Victor DeGruttola, Harvard School of Public Health

Viral genetic linkage based on data from HIV prevention trials at the community level can provide insight into HIV transmission dynamics and the impact of prevention interventions. Analysis of clustering which utilize phylogenetic methods have the potential to inform whether recently-infected individuals are infected by viruses circulating within or outside a community. In addition, they have the potential to identify characteristics of chronically infected individuals that make their viruses likely to cluster with others circulating within a community. Such clustering can be related to the potential of such individuals to contribute to the spread of the virus, either directly through transmission to their partners or indirectly through further spread of HIV from those partners. Assessment of the extent to which individual (incident or prevalent) viruses are clustered within a community will be biased if only a subset of subjects are observed, especially if that subset is not representative of the entire HIV infected population. To address this concern, we develop a multiple imputation framework in which missing sequences are imputed based on a biological model for the diversification of viral genomes. Data from a household survey conducted in a village in Botswana are used to illustrate these methods.

email: shelleyliu@fas.harvard.edu

BAYESIAN INFERENCE FOR CORRELATED BINARY DATA VIA LATENT MODELING

Deukwoo Kwon*, University of Miami
Jeesun Jung, National Institute on Alcohol Abuse and Alcoholism, National Institutes of Health
Jun-Mo Nam, National Cancer Institute, National Institutes of Health
Yi Qian, Amgen Inc.

Correlated binary data usually arise in many clinical trials. The matched-pair design is superior in power compared to a two-sample design, and is commonly used in small-sample trials. McNemar test is well-known in this setting. Several Bayesian approaches were also developed, but the implementation requires complicated computation techniques (Ghosh et al. 2000; Kateri et al. 2001; Shi et al. 2008, 2009). We propose a simplistic Bayesian approach using continuous latent model for either difference or ratio of marginal probabilities. External information can also be easily incorporated into the model. We conduct simulation study and present a real data example.

email: DKwon@med.miami.edu

THE VALIDATION OF A BETA-BINOMIAL MODEL FOR INTRA-CORRELATED BINARY DATA

Jongphil Kim*, Moffitt Cancer Center, University of South Florida
Ji-Hyun Lee, Moffitt Cancer Center, University of South Florida

The beta-binomial model accounting for the overdispersion of binary data requires an assumption that the success probability of binary data is distributed as a beta distribution. If that assumption does not hold, the inference based upon the model may be incorrect. This paper investigates beta-binomial model validation using the intra-correlated binary data which are generated without any assumption on the distribution for success probability. In addition, a nonparametric estimator for the success probability and the intraclass correlation is compared to a parametric estimator for those data.

email: Jongphil.Kim@moffitt.org

TESTING FOR HOMOGENEOUS STRATUM EFFECTS IN STRATIFIED PAIRED BINARY DATA

Dewi Rahardja*, U.S. Food and Drug Administration
Yan D. Zhao, University of Oklahoma Health Sciences Center at Tulsa

For paired binary data, the McNemar's test is widely used to test for marginal homogeneity or symmetry for a 2 by 2 contingency table. For paired categorical data with more than two categories, we can use the Bowker's test for testing the symmetry and the Stuart-Maxwell's test for testing the marginal homogeneity of a k by k ($k > 2$) contingency table. In this paper we extend the McNemar's test in another fashion by considering a series of paired binary data where the series are defined by a stratification factor. We provide a test for testing homogeneous stratum effects. For illustration, we apply our test to a cancer epidemiology study. Finally, we conduct simulations to show that our test preserves nominal Type I error level under various scenarios under the null hypothesis.

email: rahardja@gmail.com

MARGINALIZABLE CONDITIONAL MODEL FOR CLUSTERED BINARY DATA

Rui Zhang*, University of Washington
Kwun Chuen Gary Chan, University of Washington

Analysis of clustered data can provide us information on both marginal and conditional covariate effects. Marginal model specification of mean often does not require a correct specification of correlation, while conditional models admit flexible modeling of correlation structure which can lead to more efficient inference and prediction. Clearly, the two models have different strengths. In order to combine these strengths in a single analysis, we propose a marginalizable conditional model for clustered binary data such that: 1) both marginal and conditional covariate effects can be estimated from the same model and the marginal parameters have odds ratio interpretations; 2) flexible correlation structures can be modeled to improve estimation efficiency and can extend naturally to analyze higher-level clustered data. We propose a robust estimation procedure based on alternating logistic regression so that marginal parameters can be consistently estimated even when the conditional model and random effect distribution are misspecified. The estimation procedure also has much less computation burden compared to maximum likelihood estimation. Simulations and an analysis of the Madras longitudinal schizophrenia study were carried out to show the efficiency gain of the proposed method compared to Generalized Estimating Equation.

email: zhangrui1227@gmail.com

A RANDOM EFFECTS APPROACH FOR JOINT MODELING OF MULTIVARIATE LONGITUDINAL HEARING LOSS DATA ASCERTAINED AT MULTIPLE FREQUENCIES

Mulugeta Gebregziabher*, Medical University of South Carolina
Lois J. Matthews, Medical University of South Carolina
Mark A. Eckert, Medical University of South Carolina
Andrew B. Lawson, Medical University of South Carolina
Judy R. Dubno, Medical University of South Carolina

Typical hearing loss studies involve determination of hearing threshold sound pressure levels (dB) at multiple frequencies. The most common analysis of data that arise from such studies involves separate modeling of the outcomes at each frequency for each ear. However, this type of analysis ignores the correlation among the outcomes and could lead to inefficient use of data especially when one or more of the outcomes have missing data. We propose a joint modeling approach that accounts for the correlations among the responses from the multiple frequencies due to repeated measurements, clustering by ear as well as missingness. We used data from a project that motivated this study which included about 800 subjects from the South Eastern part of the US to demonstrate the proposed method.

email: gebregz@musc.edu

ROBUST ESTIMATION OF DISTRIBUTIONAL MIXED EFFECTS MODEL WITH APPLICATION TO TENDON FIBRILOGENESIS DATA

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

A new robust statistical framework is developed for comprehensive modeling of hierarchically clustered non-Gaussian distributions. It is of interest to model features of conditional distributions beyond the means (e.g. spread or skewness) or to describe the subpopulations that are viewed as approximately normally distributed. Moreover, using the robust estimation is crucial for appropriate analysis. Here, a distributional mixed effects model (DME) with conditional distributions from any parametric family is proposed as general framework for accommodating variable non-Gaussian conditional distributions and modeling their parameters as dependent on fixed and random effects. We develop a new divergence-based methodology and computational algorithm for robust joint estimation of DME model. Performances of the proposed robust joint, maximum likelihood joint and two-stage approaches are compared for analyzing fibril diameter distributions in real animal data and in simulated data with structures similar to fibril diameter distributions. Overall, numerical studies indicate superior efficiency of joint estimation as compared to previously considered two-stage approaches, and superior accuracy of robust joint estimation for contaminated data.

email: tingtingzhan@gmail.com

35. OPTIMAL TREATMENT REGIMES AND PERSONALIZED MEDICINE

PERSONALIZED MEDICINE AND ARTIFICIAL INTELLIGENCE

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

Personalized medicine is an important and active area of clinical research involving significant statistical aspects. In this talk, we describe some recent design and methodological developments in clinical trials for discovery and evaluation of personalized medicine. Statistical learning tools from artificial intelligence, including machine learning, reinforcement learning and several newer learning methods, are beginning to play increasingly important roles in these areas. We present illustrative examples in treatment of depression and cancer. The new approaches have significant potential to improve health and well being.

email: kosorok@unc.edu

ESTIMATING OPTIMAL INDIVIDUALIZED DOSING STRATEGIES

Erica EM Moodie*, McGill University
Ben Rich, McGill University
David A. Stephens, McGill University

For drugs such as warfarin, where the therapeutic window is narrow, it is important determine dosing adaptively and on an individual basis. In this talk, I will consider the application of g-estimation and Q-learning to continuous-dose treatments on simulated data in a pharmacokinetic setting where the dose allocation mechanism is known but the true outcome model is both unknown and complex.

email: erica.moodie@mcgill.ca

INTERACTIVE Q-LEARNING

Eric B. Laber*, North Carolina State University
Kristin A. Linn, North Carolina State University
Leonard A. Stefanski, North Carolina State University

Clinicians wanting to form evidence based rules for optimal treatment allocation over time have begun to estimate such rules using data collected in observational or randomized studies. Popular methods for estimating optimal sequential decision rules from data, such as Q-learning, are approximate dynamic programming algorithms that require the modeling of nonsmooth, nonmonotone transformations of the data. Unfortunately, postulating a model for the transformed data that is adequately expressive yet parsimonious is difficult, and under many simple generative models, the most

commonly employed working models---namely linear models---are seen to be misspecified. Furthermore, such estimators are nonregular making statistical inference difficult. We propose an alternative strategy for estimating optimal sequential decision rules wherein all modeling takes place before nonsmooth transformations of the data are applied. This simple interchange of modeling and transforming the data leads to high quality estimated sequential decision rules, while in many cases allowing for simplified exploratory data analysis, model building and validation, and normal limit theory. We illustrate the proposed method using data from the STAR*D study of major depressive disorder.

email: eblaber@ncsu.edu

ESTIMATING OPTIMAL TREATMENT REGIMES FROM A CLASSIFICATION PERSPECTIVE

Baqun Zhang*, Northwestern University
Anastasios A. Tsiatis, North Carolina State University
Marie Davidian, North Carolina State University
Min Zhang, University of Michigan
Eric Laber, North Carolina State University

A treatment regime is a rule that assigns a treatment, from among a set of possible treatments, to a patient based on his/her observed characteristics. Recently, robust estimators, also known as policy-search estimators, have been proposed for estimating an optimal treatment regime. However, such estimators require the parametric form of regimes to be pre-specified, which can be chosen by practical considerations like convenience and parsimony or through ad hoc preliminary analysis. In this article, we propose a novel and general framework that transforms the problem of estimating an optimal treatment regime into a classification problem wherein the optimal classifier corresponds to the optimal treatment regime. Within this framework, the parametric form of treatment regimes does not need to prespecified and instead can be identified in a data-driven way by minimizing an expected weighted misclassification error. Moreover, because classification is supervised, standard exploratory techniques can be used to choose an appropriate class of models for the treatment regime. Furthermore, this approach brings to bear the wealth of powerful classification algorithms developed in machine learning. We applied the proposed method using classification and regression trees to simulated data and data from a breast cancer clinical trial.

email: baqun.zhang@northwestern.edu

36. STATISTICAL METHODS FOR NEXT GENERATION SEQUENCE DATA ANALYSIS: A SPECIAL SESSION FOR THE ICSA JOURNAL 'STATISTICS IN BIOSCIENCES'

STATISTICAL METHODS FOR TESTING FOR RARE VARIANT EFFECTS IN NEXT GENERATION SEQUENCING ASSOCIATION STUDIES

Xihong Lin*, Harvard School of Public Health

An increasing number of large scale sequencing association studies, such as the whole exome sequencing studies, have been conducted to identify rare genetic variants associated with disease phenotypes. Testing for rare variant effects in sequencing association studies presents substantial challenges. We first provide an overview of statistical methods for testing for rare variant effects in case-control and cohort studies, and then discuss statistical methods for meta analysis of rare variant effects in sequencing association studies and family sequencing association studies. The proposed methods are evaluated using simulation studies and illustrated using data examples.

email: xlin@hsph.harvard.edu

A MODEL FOR COMBINING DE NOVO MUTATIONS AND INHERITED VARIATIONS TO IDENTIFY RISK GENES OF COMPLEX DISEASES

Xin He, Carnegie Mellon University
Kathryn Roeder*, Carnegie Mellon University

De novo mutation, a genetic mutation that neither parent possessed nor transmitted, affects risk for many diseases. A striking example is autism. Four recent whole-exome sequencing (WES) studies of 932 autism families (mother, father and affected offspring) identified six novel risk genes from a multiplicity of de novo loss-of-function (LoF) mutations and revealed that de novo LoF mutations occurred at a twofold higher rate than expected by chance. We develop statistical methods that increase the utility of de novo, mutations by incorporating additional information concerning transmitted variation. Our statistical model relates distinct types of data through a set of genetic parameters such as mutation rate and relative risk, facilitating analysis in an integrated fashion. Inference is based on an empirical Bayes strategy that allows us to borrow information across all genes to infer parameters that would be difficult to estimate from individual genes. We validate our statistical strategy using simulations mimicking rare, large-effect mutations. We found that our model substantially increases statistical power and dominates all other methods in almost all settings we examined. We illustrate our methods with WES autism data.

email: roeder@stat.cmu.edu

INTEGRATIVE ANALYSIS OF *-SEQ DATASETS FOR A COMPREHENSIVE UNDERSTANDING OF REGULATORY ROLES OF REPETITIVE REGIONS

Sunduz Keles*, University of Wisconsin, Madison
Xin Zeng, University of Wisconsin, Madison

The ENCODE projects have generated exceedingly large amounts of genomic data towards understanding how cell type specific gene expression programs are established and maintained through gene regulation. A formidable impediment to comprehensively understanding of these ENCODE data is the lack of statistical and computational methods required to identify functional elements in repetitive regions of genomes. Although next generation sequencing technologies, embraced by the ENCODE projects, are enabling interrogation of genomes in an unbiased manner, the data analysis efforts by the ENCODE projects have thus far focused on mappable regions with unique sequence contents. This is especially true for the analysis of ChIP-seq data in which all ENCODE-adapted methods discard reads that map to multiple locations (multi-reads). This is a highly critical barrier to the advancement of ENCODE data because significant fractions of complex genomes are composed of repetitive regions; strikingly, more than half of the human genome is repetitive. We present a unified statistical model for utilizing multi-reads in *-seq datasets (ChIP-, DNase-, and FAIRE-seq) with either diffused or a combination of diffused and point source enrichment patterns. Our model efficiently integrates multiple *-seq datasets and significantly advances multi-read analysis of ENCODE and related datasets.

email: keles@stat.wisc.edu

ASSOCIATION MAPPING WITH HETEROGENEOUS EFFECTS: IDENTIFYING eQTLS IN MULTIPLE TISSUES

Matthew Stephens*, University of Chicago
Timothee Flutre, University of Chicago
William Wen, University of Michigan
Jonathan Pritchard, University of Chicago

Understanding the genetic basis of variation in gene expression is now a well-established route to identifying regulatory genetic variants, and has the potential to yield important novel insights into gene regulation, and, ultimately, the biology of disease. Statistical methods for identification of eQTLs in a single tissue or cell type are now relatively mature, and several studies have shown the benefits in power obtained by the use of appropriate statistical methods, notably data normalization, robust testing procedures, and using dimension reduction techniques to control for unmeasured confounding factors. Here we consider statistical analysis methods for an important problem that until now has received less attention: combining information effectively across expression data from multiple tissues. The aims of such studies include both the identification of regulatory variants that are shared across tissues (tissue-consistent) and

that are specific to one or a few tissues (tissue-specific). We introduce some analytical methods for this problem that analyze all tissues simultaneously, taking account of the potential heterogeneity in effects among tissues. We illustrate these methods, and their potential benefits, on a dataset of Fibroblasts, LCLs and T-cells from the same set of 80 individuals (Dimas et al 2009)

email: mstephens@uchicago.edu

37. HYPOTHESIS TESTING PROBLEMS IN FUNCTIONAL DATA ANALYSIS

EMPIRICAL DYNAMICS FOR FUNCTIONAL DATA

Hans-Georg Müller*, University of California, Davis

Under weak assumptions, functional data can be described by a nonlinear first order differential equation, coupled with a stochastic drift term. A special case of interest occurs for Gaussian processes. Given a sample of functional data, one may obtain the underlying differential equation via a simple smoothing-based procedure. Diagnostics can be based on the fraction of variance that is explained by the deterministic part of the equation. A null hypothesis of interest is that the underlying dynamic system is autonomous. Learning the dynamics of longitudinal data is illustrated for human growth. The presentation is based on several joint papers with Wenwen Tao, Nicolas Verzelen and Fang Yao.

email: hgmueller@ucdavis.edu

SIMULTANEOUS CONFIDENCE BAND FOR SPARSE LONGITUDINAL REGRESSION

Shujie Ma*, University of California, Riverside
Lijian Yang, Michigan State University
Raymond Carroll, Texas A&M University

Functional data analysis has received considerable recent attention and a number of successful applications have been reported. In literature point-wise asymptotic distributions have been obtained for estimators of the functional regression mean function, but without uniform confidence bands. The fact that a simultaneous confidence band has not been established for functional data analysis is certainly not due to lack of interesting applications, but to the greater technical difficulty in formulating such bands for functional data and establishing their theoretical properties. Specifically, the strong approximation results used to establish the asymptotic confidence level in nearly all published works on confidence bands, commonly known as Hungarian embedding, are unavailable for functional data. In this talk, we provide a limit theorem for the maximum of the normalized deviations of the estimated mean functions from the true mean functions in the functional regression model via a piecewise-constant spline smoothing approach. Using this result, we construct asymptotically simultaneous confidence bands for the underlying mean function. Simulation experiments corroborate the asymptotic theory. The confidence band procedure is illustrated by analyzing CD4 cell counts of HIV infected patients.

email: shujie.ma@ucr.edu

TESTING FOR FUNCTIONAL EFFECTS

Bruce J. Swihart*, Johns Hopkins Bloomberg School of Public Health
Jeff Goldsmith, Columbia University Mailman School of Public Health
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health

The goal of our article is to provide a transparent, robust, and computationally feasible statistical approach for testing in the context of functional linear models. In particular, we are interested in testing for the necessity of functional effects against standard linear models. Our approach is to express the coefficient function in a way that reduces to a constant under the null hypothesis; thus the null model includes the average of functional predictors as a scalar covariate. The coefficient function is modeled using a penalized spline framework, and testing is performed using likelihood and restricted likelihood ratio testing (RLRT) for zero variance components in a linear mixed model framework. This problem is nonstandard because under the null hypothesis the parameter is on the boundary of the parameter space. We extend the methodology to be of use when multiple functional predictors are observed and when observations are made longitudinally. Our methods are motivated by and applied to a large longitudinal study involving diffusion tensor imaging of intracranial white matter tracts in a susceptible cohort. Relevant software is posted as an online supplement and many of the outlined methods are implemented in the R-package refund.

email: bruce.swihart@gmail.com

FUNCTIONAL MIXED EFFECTS SPECTRAL ANALYSIS

Robert T. Krafty*, University of Pittsburgh
Martica Hall, University of Pittsburgh
Wensheng Guo, University of Pennsylvania

In many experiments, time series data can be collected from multiple units and multiple time series segments can be collected from the same unit. We discuss a mixed effects Cramer spectral representation which can be used to model and test the effects of design covariates on the second-order power spectrum while accounting for potential correlations among the time series segments collected from the same unit. The transfer function is composed of a deterministic component to account for the population-average effects and a random component to account for the unit-specific deviations. The resulting log-spectrum has a functional mixed effects representation where both the fixed effects and random effects are functions in the frequency domain. A iterative smoothing spline based procedure is offered for estimation while a bootstrap procedure is developed for performing inference on the functional fixed effects.

email: krafty@pitt.edu

38. PHARMACOGENOMICS AND DRUG INTERACTIONS: STATISTICAL CHALLENGES AND OPPORTUNITIES ON THE JOURNEY TO PERSONALIZED MEDICINE

OVERVIEW OF PHARMACOGENOMICS, GENE-GENE INTERACTION, SYSTEM GENOMICS

Marylyn D. Ritchie*, The Pennsylvania State University

Pharmacogenomics is a prominent component in the drive to implement precision medicine. In the field of pharmacogenomics, a primary goal is ensuring that each patient gets the right drug at the right dose and schedule. A variety of issues arise in the design and analysis of pharmacogenomic studies. These topics will be reviewed as an introduction to the field of Pharmacogenomics. After the review, analytic approaches for complex analysis will be discussed including genome-wide gene-gene interaction approaches and integrative systems Pharmacogenomics analysis approaches. These types of approaches have been proposed and used in Pharmacogenomics to determine functionally relevant genomic markers associated with drug response.

email: marylyn.ritchie@psu.edu

INTEGRATIVE ANALYSIS APPROACHES FOR CANCER PHARMACOGENOMICS

Brooke L. Fridley*, University of Kansas Medical Center

One of the biggest challenges in the treatment of cancer is the large inter-patient variation in clinical response observed for chemotherapies. In spite of significant developments in our knowledge of cancer genomics and pharmacogenomics, numerous challenges still exist, slowing the discovery and translation of findings to the clinic. One significant challenge is the identification of relevant genomic features important in drug response. Drug response is most likely not due to a single gene but rather a complex interacting network involving genetic variation, mRNA, miRNA, DNA methylation, and external environmental factors. In cancer pharmacogenomics, this involves both germline and tumor variation. I will discuss several integrative approaches and strategies being used to determine novel pharmacogenomics hypotheses to be

followed-up in additional functional and translational studies. Such methods include: step-wise integrative analysis, gene set and pathway analysis, interaction analysis, molecular subtype analysis and functional association analysis.

email: bfridley@kumc.edu

STATISTICAL CHALLENGES IN TRANSLATIONAL BIOINFORMATICS DRUG-INTERACTION RESEARCH

Lang Li*, Indiana University, Indianapolis

Novel drug interactions can be predicted through large-scale text mining and knowledge discovery from the published literature. Using natural language processing (NLP), the key challenge is to extract drug interaction relationship through the machine learning. We propose a hybrid mixture model and tree-based approach to extract drug interaction relationship. This two-pronged approach takes advantage of both the numerical features of reported drug interaction results and the linguistic styles of presenting drug interactions. In this talk, I will discuss the concept and method of literature-based knowledge discovery in drug interaction research and data mining-based drug interaction research using large electronic medical record databases. I will discuss the pros and cons and multiple design and analyses strategies for large-scale drug interaction screening studies that use large-scale electronic medical record databases. I will illustrate these concepts in the context of a translational bioinformatics drug interaction study on myopathy, a muscle weakness adverse drug event, elucidating on both the clinical significance and the molecular pharmacology significance.

email: lali@iupui.edu

STUDY DESIGN AND ANALYSIS OF BIOMARKER AND GENOMIC CLASSIFIER VALIDATION

Cheng Cheng*, St. Jude Children's Research Hospital

Validation study of the discovered biomarkers and classifiers is an indispensable step in translating genomic findings into clinical practice. The Institute of Medicine has issued guidelines (Evolution of Translational Omics: http://www.iom.edu/Reports/2012/Evolution-of-Translational-Omics.aspx) for biomarker discovery and validation before clinical application in deciding how to treat patients, setting high bars for biomarker validation. The Test Validation Phase requires careful and rigorous consideration of the validation study design. The classification /prediction accuracy assessed in the analytical validation depends on, among many factors, the biomarkers' classification capabilities (biological) and clinical assay variation (technical). The further step using blinded samples can be retrospective or prospective and is in some way similar to a phase-II clinical trial. A typical statistical design issue to address here is determination of the sample size needed to achieve high confidence that the classifier indeed possesses the desired accuracy for clinical use. I will discuss efficient study designs for marker validation, borrowing ideas from adaptive group sequential clinical trials.

email: cheng.cheng@stjude.org

39. TRANSLATIONAL METHODS FOR STRUCTURAL IMAGING

STATISTICAL TECHNIQUES FOR THE NORMALIZATION AND SEGMENTATION OF STRUCTURAL MRI

Russell T. Shinohara*, University of Pennsylvania Perelman School of Medicine
Elizabeth M. Sweeney, Johns Hopkins Bloomberg School of Public Health
Jeff Goldsmith, Columbia University Mailman School of Public Health
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health

While computed tomography and other imaging techniques are measured in absolute units with physical meaning, magnetic resonance images are expressed in arbitrary units that are difficult to interpret and differ between study visits and subjects. Much work in the image processing literature has centered on histogram matching and other histogram mapping techniques, but little focus has been on normalizing images to have biologically interpretable units. We explore this key goal for statistical analysis and the impact of normalization on cross-sectional and longitudinal segmentation of pathology.

email: taki.shinohara@gmail.com

STATISTICAL METHODS FOR LABEL FUSION: ROBUST MULTI-ATLAS SEGMENTATION

Bennett A. Landman*, Vanderbilt University

Mapping individual variation of head and neck anatomy is essential for radiotherapy and surgical interventions. Precise localization of affected structures enables effective treatment while minimizing impact to vulnerable systems. Modern image processing techniques enable one to establish point-wise correspondence between scans of different patients using non-rigid registration, and, in theory, allow for extremely rapid labeling of medical images via label transfer (i.e., copying of labels from atlas patients to target patients). To compensate for algorithmic and anatomical mismatch, the state of the art for atlas-based segmentation is to use a multi-atlas approach in which multiple canonical patients (with labels) are registered to a target patient (without labels); statistical label fusion is used to resolve conflicts and assign labels to the target. We will discuss our recent progress (and outstanding challenges) in determining how to optimally fuse information in the form of spatial labels.

email: bennett.landman@vanderbilt.edu

IMAGING PATTERN ANALYSIS USING MACHINE LEARNING METHODS

Christos Davatzikos*, University of Pennsylvania

During the past decade there has been increased interest in the medical imaging community for advanced pattern analysis and machine learning methods, which capture complex imaging phenotypes. This interest has been amplified by the fact that many diseases and disorders, particularly in neurology and neuropsychiatry, involve spatio-temporally complex patterns that are not easily detected or quantified. One of the most important challenges in this field has been appropriate dimensionality reduction and feature extraction/selection, i.e. finding which combination of imaging features forms most discriminatory imaging patterns. We present work along these lines, by describing a joint generative-discriminative approach based on constrained non-negative matrix factorization, aiming at extracting imaging patterns of maximal discriminatory power. Applications in structural imaging of Alzheimer's Disease and in functional MRI are presented. We also give a broader overview of other applications in this field.

email: Christos.Davatzikos@uphs.upenn.edu

A SPATIALLY VARYING COEFFICIENTS MODEL FOR THE ANALYSIS OF MULTIPLE SCLEROSIS MRI DATA

Timothy D. Johnson*, University of Michigan
Thomas E. Nichols, University of Warwick
Tian Ge, University of Warwick

Multiple Sclerosis (MS) is an autoimmune disease affecting the central nervous system by disrupting nerve transmission. This disruption is caused by damage to the myelin sheath surrounding nerves that acts as an insulator. Patients with MS have a multitude of symptoms that depend on where lesions occur in the brain and/or spinal cord. Patient symptoms are rated by the Kurtzke Functional Systems (FS) scores and the paced auditory serial addition test (PASAT) score. The eight functional systems are: 1) pyramidal; 2) cerebellar; 3) brainstem; 4) sensory; 5) bowel and bladder; 6) visual; 7) cerebral; and 8) other. Of interest is whether lesion locations can be predicted using these FS and PASAT scores. We propose an autoprobit regression model with spatially varying coefficients. The data of interest are binary lesion maps. The model incorporates both spatially varying covariates as well as patient specific, non-spatially varying covariates. In contrast to most spatial applications, in which only

one realization of a process is observed, we have multiple independent realizations one from each patient. For each patient specific covariate, we derive spatial maps of these coefficients that allow us to spatially predict lesion probabilities over the brain given covariates.

email: tdjtdj@umich.edu

40. FLEXIBLE BAYESIAN MODELING

FLEXIBLE REGRESSION MODELS FOR ROC AND RISK ANALYSIS, WITH OR WITHOUT A GOLD STANDARD

Wesley O. Johnson*, University of California, Irvine
Fernando Quintana, Pontificia Universidad Catolica de Chile

A semiparametric regression model is developed for the evaluation of a continuous medical test, such as a biomarker. We focus on the scenario where a gold-standard does not exist, is too expensive, or requires an invasive surgical procedure. To compensate we incorporate covariate information by modeling the probability of disease as a function of disease covariates. In addition, we model biomarker outcomes to depend on test covariates. This allows researchers to quantify the impact of covariates on the accuracy of a test; for instance, it may be easier to diagnose a particular disease in older individuals than it is in younger ones. We further model the distributions of test outcomes for the diseased and healthy groups using flexible semiparametric classes. Notably, the resulting regression model is shown to be identifiable under mild conditions. We obtain inferences about covariate-specific test accuracy and the probability of disease for any subject, based on their disease and test covariate information. The proposed model generalizes existing ones, and a special case applies to gold-standard data. The value of the model is illustrated using simulated data and data on the age-adjusted ability of soluble epidermal growth factor receptor (sEGFR) - a ubiquitous serum protein - to serve as a biomarker of lung cancer in men.

email: wjohnson@uci.edu

A NONPARAMETRIC BAYESIAN MODEL FOR LOCAL CLUSTERING

Juhee Lee*, The Ohio State University
Peter Mueller, University of Texas, Austin
Yuan Ji, NorthShore University HealthSystem

We propose a nonparametric Bayesian local clustering (NoB-LoC) approach for heterogeneous data. The NoB-LoC model defines local clusters as blocks of a two-dimensional data matrix and produces inference about these clusters as a nested bidirectional clustering. Using protein expression data as an example, the NoB-LoC model clusters proteins (columns) into protein sets and simultaneously creates multiple partitions of samples (rows), one for each protein set. In other words, the sample partitions are nested within the protein sets. Any pair of samples might belong to the same cluster for one protein set but not for another. These local features are different from features obtained by global clustering approaches such as hierarchical clustering, which create only one partition of samples that applies for all proteins in the data set. As an added and important feature, the NoB-LoC method probabilistically excludes sets of irrelevant proteins and samples that do not meaningfully co-cluster with other proteins and samples, thus improving the inference on the clustering of the remaining proteins and samples. Inference is guided by a joint probability model for all random elements. We provide extensive examples to demonstrate the unique features of the NoB-LoC model.

email: juheelee2@gmail.com

NONPARAMETRIC TESTING OF GENETIC ASSOCIATION AND GENE-ENVIRONMENT INTERACTION THROUGH BAYESIAN RECURSIVE PARTITIONING

Li Ma*, Duke University

In genetic association studies, a central goal is to test for the dependence between the genotypes and the response, conditional on a set of covariates (e.g. environmental variables). Compared to traditional parametric methods such as the logistic regression, nonparametric methods allow greater flexibility in the underlying mode of association that can be detected. However, it is more difficult to efficiently incorporate additional covariates in a model-free setting. Some classical nonparametric tests accomplish such conditioning by dividing the marginal contingency table into conditional ones, but this approach is undesirable in many modern settings, as there are often a large number of conditional tables, most of which are sparse due to the multi-dimensionality of the genotype and covariate spaces. We introduce a Bayesian framework for nonparametrically testing genetic association given covariates that is robust to such sparsity. Moreover, under this framework one can test for gene-environment interactions through Bayesian model selection over classes of nonparametric models specified in terms of conditional independence relationships. At the core of the framework is a nonparametric prior for the retrospective genotype distribution, constructed using a randomized recursive partitioning procedure over the corresponding contingency table.

email: li.ma@duke.edu

BAYESIAN ANALYSIS OF DYNAMIC ITEM RESPONSE MODELS IN ADAPTIVE MEASUREMENT TESTING

Xiaojing Wang*, University of Connecticut
James O. Berger, Duke University
Donald S. Burdick, MetaMetrics Inc.

Item response theory models, also called latent trait models, are widely used in measurement testing to model the latent ability/trait of individuals. However, for example, in adaptive measurement testing, when there are repeated observations available for individuals through time, a dynamic structure for the latent ability/trait needs to be incorporated into the model to accommodate changes in ability. Other complications that often arise in such settings include a violation of the common assumption that test results are conditionally independent, given ability/trait and item difficulty, and that test item difficulties may be partially specified, but subject to uncertainty. Focusing on time series dichotomous response data, a new class of state space models, called Dynamic Item Response models is proposed. The models can be applied either retrospectively to the full data or on-line, in cases where real-time prediction is needed. The models are studied through simulated examples and applied to a large collection of reading test data obtained from MetaMetrics, Inc.

email: xiaojing.wang@uconn.edu

41. STATISTICAL CHALLENGES IN ALZHEIMER'S DISEASE RESEARCH

STATISTICAL CHALLENGES IN COMBINING DATA FROM DISPARATE SOURCES TO PREDICT THE PROBABILITY OF DEVELOPING COGNITIVE DEFICITS

Shane Pankratz*, Mayo Clinic

It is becoming increasingly important to identify groups of individuals who are likely to develop Alzheimer's disease (AD) before the underlying disease processes are too advanced for effective intervention. One way to achieve this is to develop models that predict AD, or even pre-clinical cognitive impairment. Many features predict future cognitive decline, and these may be used to develop risk prediction models. Ideally, one would gather all possible data from a representative collection of participants in order to build such a statistical risk model. While many of the features associated with the risk of cognitive decline are easy to obtain (e.g. age and gender), others are difficult to measure (e.g. measures from imaging modalities, biomarkers from cerebrospinal fluid), either due to high cost or low participant acceptability. Many of these difficult to obtain features are among those with greatest potential to provide insight into the risk of cognitive decline. This presentation will provide an overview of statistical methods whose use enables the development of risk prediction models when only subsets of individuals have been measured for some of the risk factors of primary interest. These methods will be illustrated using data from the Mayo Clinic Study of Aging (U01 AG06786).

email: pankratz.vernon@mayo.edu

STATISTICAL CHALLENGES IN ALZHEIMER'S DISEASE BIOMARKER AND NEUROPATHOLOGY RESEARCH

Sharon X. Xie*, University of Pennsylvania Perelman School of Medicine
Matthew T. White, Harvard Medical School

Biomarker research in Alzheimer's disease (AD) has become increasingly important because biomarkers can signal the onset of the disease before the emergence of measurable cognitive impairments. Because intervention with disease-modifying therapies for AD is likely to be most efficacious before significant neurodegeneration has occurred, there is an urgent need for biomarker-based tests that enable a more accurate and early diagnosis of AD. One of the major statistical challenges in analyzing AD biomarkers is the large measurement error due to imperfect lab conditions (e.g., antibody difference, lot-to-lot variability, storage conditions, contamination, etc.) or temporal variability. In this talk, we will demonstrate the bias in diagnostic accuracy estimates due to measurement error in the biomarker. Under the classical additive measurement error model, we will present novel approaches to correct the bias in diagnostic accuracy estimates for a single error-prone biomarker and for two correlated error-prone biomarkers. We will then discuss future directions and challenges of AD biomarker research with specific examples and motivations. Finally, postmortem examination of the underlying cause of dementia is important in AD research. Statistical challenges in how to analyze neuropathology data will be discussed.

email: sxie@mail.med.upenn.edu

FUNCTIONAL REGRESSION FOR BRAIN IMAGING

Xuejing Wang, University of Michigan
Bin Nan*, University of Michigan
Ji Zhu, University of Michigan
Robert Koeppe, University of Michigan

It is well-known that the major challenges in analyzing imaging data are from spatial correlation and high-dimensionality of voxels. Our primary motivation and application come from brain imaging studies on cognitive impairment in elderly subjects with brain disorders. We propose an efficient regularized Haar wavelet-based approach for the analysis of three-dimensional brain image data in the framework of functional data analysis, which automatically takes into account the spatial information among neighboring voxels. We conduct extensive simulation studies to evaluate the prediction performance of the proposed approach and its ability to identify related regions to response variable, with the underlying assumption that only few relatively small subregions are associated with the response variable. We then apply the proposed method to searching for brain subregions that are associated with cognition using PET images of patients with Alzheimer's disease, patients with mild cognitive impairment, and normal controls. Additional challenges, current and future directions of statistical methods in imaging analysis of AD will also be discussed.

email: bnan@umich.edu

STATISTICAL CHALLENGES IN ALZHEIMER'S DISEASE CLINICAL TRIAL AND EPIDEMIOLOGIC RESEARCH

Steven D. Edland*, University of California, San Diego

Alzheimer researchers are actively pursuing clinical trails and cohort studies targeting the earliest stages of disease, before clinically diagnosable disease has manifest. Changes in cognitive and functional symptoms, the typical outcome variables in these studies, are difficult to detect at this earliest stage of disease. This can be address by modifying study design (e.g. enrichment strategies), considering alternative surrogate biomarker endpoints, or using statistical methods such as item response theory to optimize the performance of available measures as endpoints. This talk will review these approaches, describe statistical methods for comparing the relative efficiency of different approaches, and discuss potential new directions for this area of active statistical research.

email: sedland@ucsd.edu

42. DIAGNOSTIC AND SCREENING TESTS

MISSING DATA EVALUATION IN DIAGNOSTIC MEDICAL IMAGING STUDIES

Jingjing Ye*, U.S. Food and Drug Administration
Norberto Pantoja-Galicia, U.S. Food and Drug Administration
Gene Pennello, U.S. Food and Drug Administration

When evaluating the accuracy of a new diagnostic medical imaging modality, multi-reader multi-case (MRMC) studies are frequently used to compare the diagnostic performance of the new modality with a standard modality. Missing data, including missing verification of disease state and/or reader determination on the cases are often encountered in MRMC studies. For example, missing data can occur when patients are lost-to-follow-up to confirm them as non-cancer cases, or patients are excluded from random selection because of quality control problems. Naïve estimates based on completer analyses may introduce bias in diagnostic accuracy estimates. In this talk, we will review the type and possible scenarios of missing data in MRMC studies. In addition, several analyses methods based on assumptions of Missing at Random (MAR) or Not Missing at Random (NMAR) will be introduced and compared. In particular, the conservative method of non-informative imputation (NI) will be defined and applied. Additionally a tipping-point analysis is proposed to be applied to evaluate missing data in MRMC settings.

email: jingjingye@gmail.com

ESTIMATING WEIGHTED KAPPA UNDER TWO STUDY DESIGNS

Nicole Blackman*, CSL Behring

Weighted kappa may be used as a measures of association between an experimental test and a gold standard test. Statistical properties of estimators of weighted kappa are evaluated for cross-sectional and case-control sampling. The aims of this investigation are: to quantify the benefits of case-control versus cross-sectional sampling; to determine the effects of prevalence, and the effects of the relative gravity of a false negative versus a false positive discrepancy; and to determine how well sampling variability is estimated. Implications of the findings for design and analysis of medical test evaluation studies are addressed.

email: Nicole.Blackman@TheoremClinical.com

EFFICIENT POOLING METHODS FOR SKEWED BIOMARKER DATA SUBJECT TO REGRESSION ANALYSIS

Emily M. Mitchell*, Emory University Rollins School of Public Health
Robert H. Lyles, Emory University Rollins School of Public Health
Michelle Danaher, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Neil J. Perkins, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Enrique F. Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Pooling biological specimens prior to performing expensive laboratory tests can considerably reduce costs associated with certain epidemiologic studies. Recent research illustrates the potential for minimal information loss when efficient pooling strategies are performed on an outcome in a linear regression setting. Many public health studies, however, involve skewed data, often assumed to be log-normal or gamma distributed, which complicates the regression estimation procedure. We use simulation studies to assess various analytical methods for performing generalized regression on a skewed, pooled outcome, including the application of a Monte Carlo EM algorithm. The potential consequences of distributional misspecification as well as various methods to identify and apply the appropriate assumptions based on the pooled data are also discussed.

email: emitch8@emory.edu

ADAPTING SHANNON'S ENTROPY TO STRENGTHEN RELATIONSHIP BETWEEN TIME SINCE HIV-1 INFECTION AND WITHIN-HOST VIRAL DIVERSITY

Natalie M. Exner*, Harvard University
Marcello Pagano, Harvard University

Within-host viral diversity is a potential predictor of time since infection with HIV-1. A standard measure of viral diversity is normalized Shannon's entropy using viral sequence patterns. We argue that this version of Shannon's entropy is not well-suited for predicting time since infection because it can reach a maximal value even when sequences are highly similar. Furthermore, if an individual is infected by multiple viruses, the relationship between time since infection and diversity is confounded. We propose two modifications to improve the utility of within-host viral diversity as a predictor of time since infection: (1) divide alignment into sections of length L, evaluate entropy in each section, and combine information into an overall score; (2) assess each sample for multiplicity of infection, and, if multiply infected, entropy is measured in separate lineages and combined into an overall score. The methods were compared by fitting generalized estimating equations (GEE) models with diversity as the independent variable predicting time since seroconversion. The optimal L was approximately 200 nucleotides. By making simple adjustments to the measure of viral diversity, we strengthened the relationship with time since infection as evidenced by increases in the GEE Wald test statistics.

email: nexner@hsph.harvard.edu

ON USE OF PARTIAL AREA UNDER THE ROC CURVE FOR EVALUATION OF DIAGNOSTIC PERFORMANCE

Hua Ma*, University of Pittsburgh
Andriy I. Bandos, University of Pittsburgh
Howard E. Rockette, University of Pittsburgh
David Gur, University of Pittsburgh

Accuracy assessment is important in many fields including diagnostic medicine. The methodology for statistical analysis of diagnostic performance continues to develop to improve the efficiency and practically relevance of the inferences. This article focuses on the partial area under the receiver operating characteristic (ROC) curve, or pAUC, which is often more clinically relevant than the area under the entire ROC curve. Despite its relevance the pAUC is not used frequently because of the apprehended statistical limitations. The partial area index and the standardized pAUC emerged to remedy some of these limitations, however, other problems remain. We derive two important properties of the standardized pAUC which could facilitate a wider and more appropriate use of this important summary index. First, we prove that standardized pAUC increases with increasing range for practically common ROC curves. Second, we demonstrate that, contrary to common belief, the variance of the standardized pAUC can decrease with increasing range. Our results indicate that increasing range will likely increase the standardized pAUC, thus the standardization does not eliminate the need to consider the range in the interpretation of the results. Futhermore, under many scenarios pAUC offers more efficient inferences than the area under the entire ROC curve.

email: hum9@pitt.edu

MULTIREADER ANALYSIS OF FROC CURVES

Andriy Bandos*, University of Pittsburgh

Many contemporary problems of accuracy assessment must deal with detection and localization of multiple targets per subject. In diagnostic medicine new systems are frequently evaluated with respect to their accuracy of detection and localization of possibly multiple abnormal lesions per patient. Assessment of detection-localization accuracy requires extensions of the conventional ROC analysis, one of most known of which is the free-response ROC (FROC) analysis. The FROC curve is a fundamental summary of detection-localization performance and several of its summary indices gave rise to performance-assessment methods for a standalone diagnostic system. Many diagnostic systems require interpretation by a trained human observer and therefore are frequently evaluated in the fully-crossed multireader studies. Analysis of the multireader FROC studies is non-trivial and straightforward extensions of the existing multireader techniques were known to fail for some summary indices. We propose a closed-form method for multireader analysis using the existing summary index for the FROC curve. The method is based on analytical reduction of bias of the ideal bootstrap variance of the reader-averaged summary index. We present results of the simulation study demonstrating the improved efficiency of the proposed method as compared to the non-parametric bootstrap approach.

email: anb61@pitt.edu

COMPARING TWO CORRELATED C INDICES WITH RIGHT-CENSORED SURVIVAL OUTCOME: A NONPARAMETRIC APPROACH

Le Kang*, U.S. Food and Drug Administration
Weijie Chen, U.S. Food and Drug Administration
Nicholas Petrick, U.S. Food and Drug Administration

The area under the receiver operating characteristic (ROC) curve (AUC) is often used as a summary index of diagnostic ability, although its use is limited in evaluating biomarkers with censored survival outcome. The overall C index, motivated as an extension of AUC viewed through the Mann-Whitney-Wilcoxon U statistic, has been proposed by Harrell as a concordance measure between right-censored survival outcome and a predictive biomarker. Asymptotic variance estimations for C as well as associated confidence intervals have been investigated (e.g., Pencina and D'Agostino). To our knowledge, there is no published result on the comparison of two correlated C indices resulting from two biomarkers or two algorithms combining multiple biomarkers into composite patient scores. In this work, we develop estimators for the variance of the difference between two correlated C indices as well as the statistic for hypothesis testing. Our approach utilizes results for Kendall's tau together with the delta method. We evaluate the performance of the statistic in terms of the type I error rate and power via simulation studies, and we compare this performance to that of a bootstrap resampling approach. Our preliminary result shows that the proposed statistic has satisfactory type I error control.

email: Le.Kang@fda.hhs.gov

43. CAUSAL INFERENCE AND COMPETING RISKS

ESTIMATION OF VACCINE EFFECTS USING SOCIAL NETWORK DATA

Elizabeth L. Ogburn*, Harvard University
Tyler J. VanderWeele, Harvard University

There is a growing literature on the possibility of testing for the presence of different causal mechanisms in social networks and a consensus that more rigorous methods are needed. We demonstrate the possibility of testing for the presence of two different indirect effects of vaccination on an infectious disease outcome using data from a single social network. The indirect effect of one individual's vaccination on another's outcome can be decomposed into the infectiousness and contagion effects, representing two distinct causal pathways by which one person's vaccination may affect another's disease status. The contagion effect is the indirect effect that vaccinating one individual may have on another by preventing the vaccinated individual from getting the disease and thereby from passing it on. The infectiousness effect is the indirect effect that vaccination might have if, instead of preventing the vaccinated individual from getting the disease, it renders the disease less infectious, thereby reducing the probability that the vaccinated infected individual transmits the disease. For heuristic purposes we will focus on the paradigmatic example of the effect of a vaccination on an infectious disease outcome; however, effects like the contagion and infectiousness are of interest in other settings as well.

email: eogburn@hsph.harvard.edu

ESTIMATING COVARIATE EFFECTS BY TREATING COMPETING RISKS

Bo Fu*, University of Pittsburgh
Chung-Chou H. Chang, University of Pittsburgh

In analyses of time-to-event data from clinical trials or observational studies, it is important to account for informative dropouts that are due to competing risks. If researchers fail to account for the association between the event of interest and informative dropouts, they may encounter unknown amplitude bias when they identify the effects of potential risk factors related to time to the main cause of failure. In this article, we propose an approach that jointly models time to the main event and time to the competing events. The approach uses a set of random terms to capture the dependence between the main and competing events. It offers two fundamental likelihood functions that have different structures for the random terms but may be combined in practice. To estimate the unknown covariate effects by optimizing the joint likelihood functions, we used three methods: the Gaussian quadrature method, the Bayesian-Markov chain Monte Carlo method, and the hierarchical likelihood method. We compared the performance of these methods via simulations and then applied them to identify risk factors for Alzheimer's disease and other forms of dementia.

email: bof5@pitt.edu

IDENTIFIABILITY OF MASKING PROBABILITIES IN THE COMPETING RISKS MODEL

Ye Liang*, Oklahoma State University
Dongchu Sun, University of Missouri

Masked failure data arise in both reliability engineering and epidemiology. The phenomenon of masking occurs when a subject is exposed to multiple risks. A failure of the subject can be caused by one of the risks, but the cause is unknown or known up to a subset of all risks. In reliability engineering, a device may fail because of one of its defective components. However, the precise failure cause is often unknown due to lack of proper diagnostic equipment or prohibitive costs. In epidemiology, sometimes the cause of death for a patient is not known exactly due to missing or partial information on the state death certificate. A competing risks model with masking probabilities is widely used for the masked failure data. However, in many cases, the model suffers from an identification problem. Without proper restrictions, the masking probabilities in the model could be nonestimable. Our work reveals that the identifiability of masking probabilities depends on both the masking structure of data and the cause-specific hazard functions. Motivated by this result, existing solutions are reviewed and further improved. The improved solutions aim to achieve minimum compromises, and thus are cost-efficient in practices. A Bayesian framework is adopted and discussed in the statistical inference.

email: ye.liang@okstate.edu

ADJUSTING FOR OBSERVATIONAL SECONDARY TREATMENTS IN ESTIMATING THE EFFECTS OF RANDOMIZED TREATMENTS

Min Zhang*, University of Michigan
Yanping Wang, Eli Lilly and Company

In randomized clinical trials, for example, on cancer patients, it is not uncommon that patients may voluntarily initiate a secondary treatment post randomization, which needs to be properly adjusted for in estimating the true effects of randomized treatments. As an alternative to the approach based on a marginal structural Cox model (MSCM) in Zhang and Wang (2012), we propose methods that view the time to start a secondary treatment as a dependent censoring process, which is handled separately from the usual censoring such as loss to follow-up. Two estimators are proposed, both based on the idea of inverse weighting by the probability of having not started a secondary treatment yet, and the second estimator focuses on improving efficiency of inference by a robust covariate-adjustment that does not require any additional assumptions. The proposed methods are evaluated and compared with the MSCM-based method in terms of bias and variance tradeoff using simulations and application to a cancer clinical trial.

email: mzhangst@umich.edu

COMPARING CUMULATIVE INCIDENCE FUNCTIONS BETWEEN NON-RANDOMIZED GROUPS THROUGH DIRECT STANDARDIZATION

Ludi Fan*, University of Michigan
Douglas E. Schaubel, University of Michigan

Competing risks data arise naturally in many biomedical studies since the subject is often at risk for one of many types of events that would preclude him/her from experiencing all other events. It is often of interest to compare outcomes between subgroups of subjects. In the presence of observational data, group is typically not randomized, so that adjustment must be made for differences in covariate distributions across groups. The proposed method aims to compare the cumulative incidence function (CIF) between subgroups of subjects from an observational study by a measure based on direct standardization that contrasts the population average cumulative incidence under two scenarios: (i) subjects are distributed across groups as per the existing population (ii) all subjects are members of a particular group. The proposed comparison of CIFs has a strong connection to measures used in the causal inference literature. The proposed methods are semi-parametric in the sense that no models are assumed for the cause-specific hazards or the subdistribution function. Observed event counts are weighted using Inverse Probability of Treatment Weighting (IPTW) and Inverse Probability of Censoring Weighting (IPCW). We apply the proposed method to national kidney transplantation data from the Scientific Registry of Transplant Recipients (SRTR).

email: lfan@umich.edu

IMPROVING MEDIATION ANALYSIS BASED ON PROPENSITY SCORES

Yeying Zhu*, The Pennsylvania State University
Debashis Ghosh, The Pennsylvania State University
Donna L. Coffman, The Pennsylvania State University

In mediation analysis, researchers are interested in examining whether a randomized treatment or intervention may affect the outcome through an intermediate factor. Traditional mediation analysis (Baron and Kenny, 1986) applies a structure equation modeling (SEM) approach and decomposes the intent-to-treat (ITT) effect into direct and indirect effects. More recent approaches interpret the mediation effects based on potential outcome framework. In practice, there often exist confounders, pre-treatment covariates that jointly influence the mediator and the outcome. Under the sequential ignorability assumption, propensity-score-based methods are often used to adjust for confounding and reduce the dimensionality of confounders simultaneously. In this article, we show that combining machine learning algorithms (such as a generalized boosting model) and logistic regression to estimate propensity scores can be more accurate and efficient in estimating the controlled direct effects, compared to logistic regression only. The proposed methods are general in the sense that we can combine multiple candidate models to estimate propensity scores and use the cross-validation criterion to select the optimal subset of the candidate models for combining.

email: yxz165@psu.edu

ON THE NONIDENTIFIABILITY PROPERTY OF ARCHIMEDEAN COPULA MODELS

Antai Wang*, Columbia University

In this talk, we present a peculiar property shared by the Archimedean copula models, that is, different Archimedean copula models with distinct dependent levels can have the same crude survival functions for dependent censored data. This property directly shows the nonidentifiability property of the Archimedean copula models. The proposed procedure is then demonstrated by two examples.

email: aw2644@columbia.edu

44. EPIDEMIOLOGIC METHODS AND STUDY DESIGN

CALIBRATING SENSITIVITY ANALYSIS TO OBSERVED COVARIATES IN OBSERVATIONAL STUDIES

Jesse Y. Hsu*, University of Pennsylvania
Dylan S. Small, University of Pennsylvania

In medical sciences, statistical analyses based on observational studies are common phenomena. One peril of drawing inferences about the effect of a treatment on subjects using observational studies is the lack of randomized assignment of subjects to the treatment. After adjusting for measured pretreatment covariates, perhaps by matching, a sensitivity analysis examines the impact of an unobserved covariate, u, in an observational study. One type of sensitivity analysis uses two sensitivity parameters to measure the degree of departure of an observational study from randomized assignment. One sensitivity parameter relates u to treatment and the other relates u to response. For subject matter experts, it may be difficult to specify plausible ranges of values for the sensitivity parameters on their absolute scales. We propose an approach that calibrates the values of the sensitivity parameters to the observed covariates and is more interpretable to subject matter experts. We will illustrate our method using data from the U.S. National Health and Nutrition Examination Survey regarding the relationship between cigarette smoking and blood lead levels.

email: hsu9@wharton.upenn.edu

OPTIMAL FREQUENCY OF DATA COLLECTIONS FOR ESTIMATING TRANSITION RATES IN A CONTINUOUS TIME MARKOV CHAIN

Chih-Hsien Wu*, University of Texas School of Public Health

A finite-state continuous time Markov chain model is often used to describe changes of disease stages. The transition rates are often used to characterize the process and predict the future behavior of the process. In practice, since the outcome data are often observed periodically, times of stage changes and the duration that a process stays in one stage are usually not directly observable. There could be more than one change occurred between two observational instances, or no change occurred for several observed time periods. The former may lead to bias in estimation and the later may consume inadequate study resources. In this study, we propose a simulation study to examine the optimal data-collection frequency under the constraint of a fixed study cost or a target statistical power.

email: Chih-Hsien.Wu@uth.tmc.edu

SOURCE-SINK RECONSTRUCTION THROUGH REGULARIZED MULTI-COMPONENT REGRESSION ANALYSIS

Kun Chen*, Kansas State University
Kung-Sik Chan, University of Iowa

The problem of reconstructing the source-sink dynamics arises in many biological systems. Our research is motivated by marine applications where newborns are passively dispersed by ocean currents from several potential spawning sources to settle in various nursery regions that collectively constitute the sink. The reconstruction of the source-sink linkage pattern, i.e., to identify which sources contribute to which regions in the sink, is a foremost and challenging task in marine ecology. We derive a nonlinear multi-component regression model for source-sink reconstruction, which is capable of simultaneously selecting important linkages from the sources to the sink regions and making inference about the unobserved spawning activities at the sources. A sparsity-inducing and nonnegativity-constrained regularization approach is developed for model estimation, and theoretically we show that our method enjoys the oracle properties. We apply the proposed approach to study the observed jellyfish habitat expansion in the East Bering Sea. It appears that the jellyfish habitat expansion resulted from the combined effects of higher jellyfish productivity due to warmer climate and wider circulation of the jellyfish owing to a shift in the ocean circulation pattern starting in 1990.

email: kunchen@ksu.edu

REGRESSION MODELS FOR GROUP TESTING DATA WITH POOL DILUTION EFFECTS

Christopher S. McMahan, Clemson University
Joshua M. Tebbs*, University of South Carolina, Columbia
Christopher R. Bilder, University of Nebraska, Lincoln

Group testing is widely used to reduce the cost of screening individuals for infectious diseases. There is an extensive literature on group testing, most of which traditionally has focused on estimating the probability of infection in a homogeneous population. More recently, this research area has shifted towards estimating individual-specific probabilities in a regression context. However, existing regression approaches have assumed that the sensitivity and specificity of pooled biospecimens are constant and do not depend on the pool sizes. For those applications where this assumption may not be realistic, these existing approaches can lead to inaccurate inference, especially when pool sizes are large. Our new approach, which exploits the information readily available from underlying continuous biomarker distributions, provides reliable inference in settings where pooling would be most beneficial and does so even for larger pool sizes. We illustrate our methodology using hepatitis B data from a study involving Irish prisoners.

email: tebbs@stat.sc.edu

BAYESIAN ADJUSTMENT FOR CONFOUNDING IN THE PRESENCE OF MULTIPLE EXPOSURES

Krista Watts*, Harvard School of Public Health
Corwin M. Zigler, Harvard School of Public Health
Chi Wang, University of Kentucky
Francesca Dominici, Harvard School of Public Health

While currently most epidemiological studies examine health effects associated with exposure to a single environmental contaminant at a time, there has been a shift of interest to multiple pollutants. We propose an extension to Bayesian Adjustment for Confounding (Wang et al., 2012) to estimate the joint effect of a simultaneous change in more than one exposure while addressing uncertainty in the selection of the confounders. Our approach is based on specifying models for each exposure and the outcome. We perform Bayesian variable selection on all models and link them through our specification prior odds of including a predictor in the outcome model, given its inclusion in the exposure models. In simulation studies we show that our method, BAC for multiple exposures (BAC-ME) estimates the joint effect with smaller bias and mean squared error than traditional Bayesian Model Averaging (BMA) or adaptive LASSO. We compare these methods in a cross-sectional data set of hospital admissions, air pollution levels, and weather and census variables. Using each approach, we estimate the change in emergency hospital admissions associated with a simultaneous change in PM2.5 and ozone from their average levels to the National Ambient Air Quality Standards (NAAQS) set by the EPA.

email: kwatts@hsph.harvard.edu

EXPLORING THE ADDED VALUE OF IMPOSING AN OZONE EFFECT MONOTONICITY CONSTRAINT AND OF JOINTLY MODELING OZONE AND TEMPERATURE EFFECTS IN AN EPIDEMIOLOGIC STUDY OF AIR POLLUTION AND MORTALITY

James L. Crooks*, U.S. Environmental Protection Agency
Lucas Neas, U.S. Environmental Protection Agency
Ana G. Rappold, U.S. Environmental Protection Agency

Epidemiologic studies have shown that both ozone and temperature are associated with increased risk for cardio-respiratory mortality and morbidity. The current study seeks to understand the impact on the risk surface of jointly modeling the nonlinear effects of ozone and temperature, and of imposing positive monotonicity on the ozone risk. To this end, a flexible Bayesian model was developed. The independent, nonlinear effects of temperature and ozone on mortality are modeled using M- and I-spline basis functions, and outer products of these functions are used to model the joint effect. Positive monotonicity on the ozone risk rate is imposed by placing a prior distribution on the I-spline and the I-spline/B-spline coefficients that has a drop-off below

zero. The hyper-parameter controlling the shape of the drop-off can be chosen to soften or harden the threshold as desired. The model was applied to the data from the US National Morbidity, Mortality, and Air pollution Study for 95 major US urban centers between 1987 and 2000. Results are compared to those obtained under the assumption of linear effect of ozone without positivity constraint (Bell et al, JAMA 2004). [Disclaimer: This work does not reflect official EPA policy].

email: jimcrooks1975@gmail.com

STABLE MODEL CONSTRUCTION USING FRACTIONAL POLYNOMIALS OF CONTINUOUS COVARIATES FOR POISSON REGRESSION WITH APPLICATION TO LINKED PUBLIC HEALTH DATA

Michael Regier*, West Virginia University
Ruoxin Zhang, West Virginia University

Fractional polynomials provide a method by which we can consider a wide range of non-linear relationships between the outcome and predictors. There is active research in the use of fractional polynomials for linear, logistic and survival models, but investigations into their use for Poisson regression models are limited. In this work, we use a Poisson fractional polynomial model as an interpretable smoother with application to a linked public health data set to investigate the ecological relationship between county water lithium levels and suicide rates using the linked data from the Texas Department of State Health, the US Census Bureau, and the Texas Water Development Board Groundwater Database. We compare fractional polynomials and splines for modeling rates, investigate the effect of influential points, and discuss model interpretation. Finally, we propose a method for stable model construction and compare this new proposal to the current multivariable fractional polynomial selection algorithm.

email: mregier@hsc.wvu.edu

45. LONGITUDINAL DATA: METHODS AND MODEL SELECTION

AIC-TYPE MODEL SELECTION CRITERION FOR LONGITUDINAL DATA INCORPORATING GEE APPROACH

Hui Yang*, University of Rochester Medical Center
Guohua Zou, Chinese Academy of Sciences
Hua Liang, University of Rochester Medical Center

Akaike Information Criterion, which is based on maximum likelihood estimation and cannot be applied directly to the situations when likelihood functions are not available, has been modified for model selection in longitudinal data with generalized estimating equations via working independence model. This paper proposes another modification to AIC, the difference between the quasi-likelihood of a candidate model and of a narrow model plus a penalty term. Such difference avoids calculating complex integration from quasi-likelihood and inherits large sample behaviors from AIC. As a by-product, we also give a theoretical justification of the equivalence in distribution between quasi-likelihood ratio test and Wald test incorporating GEE approach. Numerical performance supports its preference to its competitors. The proposed criterion is applied to analyze a real dataset as an illustration.

email: leochrishy@gmail.com

VARIABLE SELECTION FOR FAILURE TIME DATA FROM STRATIFIED CASE-COHORT STUDIES: AN APPLICATION TO A RETROSPECTIVE DENTAL STUDY

Sangwook Kang*, University of Connecticut
Cheolwoo Park, University of Georgia
Daniel J. Caplan, University of Iowa
Young joo Yoon, Daejeon University

Root canal therapy (RCT) is an option to extend the life of a damaged tooth. But even after RCT, teeth still can be lost due to many reasons. In a retrospective dental study, it was of interest to identify factors associated to survival of root canal filled (RCF) teeth. To accomplish this goal, we propose a variable selection procedure for Cox models via a penalization of estimating functions. The proposed method is developed under a stratified case-cohort design, which includes the case-control design considered in the dental study as a special case. An adaptive bridge method along with other penalized regression methods are adapted to appropriately accommodate this study design. Both asymptotic properties and finite sample properties of the proposed methods are investigated. We illustrate our methods to the retrospective dental study data.

email: sangwook.kang@uconn.edu

CORRECTING THE EFFECTS OF MODEL SELECTION IN LINEAR-MIXED EFFECT MODELS

Adam P. Sima*, Virginia Commonwealth University

In linear regression models it has been well documented that, after model selection has occurred, the ordinary least squares estimate of the variance is biased downward and coverage probabilities are often less than nominal size. Less discussed is how model selection affects the variance estimates and coverage probabilities when the response variable may not be considered independent. Through a simulation study, it is shown that the problems that are present in the linear regression models are present in linear mixed-effects models. An estimate of the variance based on the concept of generalized degrees of freedom is presented that reduces the downward bias. Furthermore, methodology is presented that inflates the covariance matrix of the fixed effect parameters so that confidence intervals have close to nominal size. The application of this novel methodology is presented in the context of a clinical trial for treatment of cervical radiculopathy.

email: simaa@vcu.edu

A RANDOM-EFFECTS MODEL FOR LONGITUDINAL DATA WITH A RANDOM CHANGE-POINT AND NO TIME ZERO: AN APPLICATION TO MODELING AND PREDICTION OF INDIVIDUALIZED LABOR CURVES

Paul Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Alexander C. McLain*, University of South Carolina

In some longitudinal studies the initiation time of the process is not clearly defined, yet it is important to make inference or do predictions about the longitudinal process. The application of interest in this paper is to provide a framework for modeling individualized labor curves (longitudinal cervical dilation measurements) where the start of labor is not clearly defined. This is a well-known problem in obstetrics where the benchmark reference time is often chosen as the end of the process (individuals are fully dilated at 10cm) and time is run backwards. It is of interest to develop a model that can use past dilation measurements to predict future observations to aid obstetricians in determining if a labor is on a suitable trajectory. The backwards time framework is not applicable to our problem, since a main interest of this project is providing dynamic individualized predictions of the longitudinal curve (where backwards time is unknown). We propose to model the longitudinal labor dilations with a random-effects model with unknown time-zero and a random change point. We present a MC-EM approach for parameter estimation as well as for dynamic prediction of the future curve from past dilation measurements. The methodology is illustrated with longitudinal cervical dilation data from the Consortium of Safe Labor Study.

email: mclaina@mailbox.sc.edu

PARAMETER ESTIMATION FOR HIV DYNAMIC MODELS INCORPORATING LONGITUDINAL STRUCTURE

Yao Yu*, University of Rochester
Hua Liang, University of Rochester

We apply nonlinear mixed-effects models (NLME) to estimate the fixed-effects and random-effects for dynamic parameters in Perelson's HIV dynamic model. The estimator maintains biological interpretability for dynamic parameters because this method is based on numerical solutions of ODEs rather than close-form solutions. This approach is applied to a real data set collected from an AIDS clinical trial. The real data analysis is integrated with simulation studies. Eight different settings are used to verify the reliability of the obtained estimates and explore possible approaches to improve the accuracy of the estimators. Our simulation results demonstrate small biases of the fixed-effects estimates in settings where the individual observations are rich as well as sparse. For random-effects, settings with small measurement errors or large sample size provide us good estimates of variance components, which indicate the effectiveness of this approach.

email: Yao_Yu@urmc.rochester.edu

TWO-STEP SMOOTHING ESTIMATION OF CONDITIONAL DISTRIBUTION FUNCTIONS BY TIME-VARYING PARAMETRIC MODELS FOR LONGITUDINAL DATA

Mohammed R. Chowdhury*, The George Washington University
Colin O. Wu, National Heart, Lung and Blood Institute, National Institutes of Health
Reza Modarres, The George Washington University

Nonparametric estimation of conditional CDFs with longitudinal data have important applications in biomedical studies. A class of two-step nonparametric estimators of the conditional CDFs with longitudinal data have been proposed in Wu and Tian (2012) without assuming any structures for the model. This unstructured nonparametric approach may not lead to adequate estimators when y(t) is close to the boundary of the support. We study a structural nonparametric approach for estimating the conditional CDF with a longitudinal sample. Our estimation method relies on a two-step smoothing method, in which we first estimate the time-varying conditional CDF at a set of time points, and then use the local polynomial method or the kernel method to compute the smooth estimators of CDF based on the raw estimators. All asymptotic properties and asymptotic distribution of our estimators have been derived. Applications of our two-step estimation method have been demonstrated using the NGHS blood pressure data. Finite sample properties of these local polynomial estimators are investigated and compared through a simulation study with a longitudinal design similar to the NGHS. Our simulation results indicate that our CDF estimators based on the time-varying parametric models may be superior to the unstructured nonparametric estimators, particularly when y(t) is near the boundary of the support.

email: mohammed@gwmail.gwu.edu

FIDUCIAL GENERALIZED p-VALUES FOR TESTING ZERO-VARIANCE COMPONENTS IN LINEAR MIXED-EFFECTS MODELS

Haiyan Su*, Montclair State University
Xinmin Li, Shandong University of Technology
Hua Liang, University of Rochester
Hulin Wu, University of Rochester

Linear mixed-effects models are widely used in analysis of longitudinal data. However, testing for zero-variance components of random effects has not been well resolved in statistical literature, although some likelihood-based procedures have been proposed and studied. In this article, we propose a generalized p-value based method in coupling with fiducial inference to tackle this problem. The proposed method is also applied to test linearity of the nonparametric functions in additive models. We provide theoretical justifications and develop an implementation algorithm for the proposed method. We evaluate its finite-sample performance and compare it with that of the restricted likelihood ratio test via simulation experiments. An application of real study using the proposed method is also provided.

email: suh@mail.montclair.edu

46. SPATIAL/TEMPORAL MODELING

A COMBINED ESTIMATING FUNCTION APPROACH FOR FITTING STATIONARY POINT PROCESS MODELS

Chong Deng*, Yale University
Rasmus P. Waagepetersen, Aalborg University, Denmark
Yongtao Guan, University of Miami

The composite likelihood estimation (CLE) method is a computationally simple approach for fitting spatial point process models. The construction of the CLE inherently assumes that different pairs of events are independent. Such an assumption is invalid and may lead to efficiency loss in the resulting estimator. We propose a new estimation procedure that improves the efficiency. Our approach is to 'optimally' combine two sets of estimating functions, where one set is derived from the CLE while the other is related to the correlation among the different pairs of events. Our proposed method can be used to fit parametric models from a variety of spatial point processes including both Poisson cluster and log Gaussian Cox processes. We demonstrate its efficacy through a simulation study and an application to the longleaf pine data.

email: chong.deng@yale.edu

CHILDHOOD CANCER RATES, RISK FACTORS, & CLUSTERS: SPATIAL POINT PROCESS APPROACH

Md M. Hossain*, Cincinnati Children's Hospital Medical Center

Childhood cancer incidences for the state of Ohio for diagnosis year: 1996-2009 was analyzed by using the spatial point process and the area level modeling approaches. Effects for the environmental, agricultural, and demographic risk factors observed at various levels are adjusted by using the multilevel modeling framework. Results from this ongoing research will be presented. The issues involved with multivariate point process by considering various cancer sites (e.g., leukemias, brain tumors or lymphomas) jointly will also be addressed.

email: md.hossain@cchmc.org

BRIDGING CONDITIONAL AND MARGINAL INFERENCE FOR SPATIALLY-REFERENCED BINARY DATA

Laura F. Boehm*, North Carolina State University
Brian J. Reich, North Carolina State University
Dipankar Bandyopadhyay, University of Minnesota

Spatially-referenced binary data are common in epidemiology and public health. Owing to its elegant log-odds interpretation of the regression coefficients, a natural model for these data is logistic regression. However, to account for missing confounding variables that might exhibit a spatial pattern (say, socioeconomic, biological or environmental conditions), it is customary to include a Gaussian spatial random effect. Conditioned on the spatial random effect, the coefficients may be interpreted as log odds ratios. However, marginally over the random effects, the coefficients no longer preserve the log-odds interpretation, and the estimates are hard to interpret and generalize to other spatial regions. To resolve this issue, we propose a new spatial random effect distribution through a copula framework which ensures that the regression coefficients maintain the log-odds interpretation both conditional on and marginally over the spatial random effects. We present simulations to assess the robustness of our proposition to various random effects, and apply it to an interesting dataset assessing periodontal health of Gullah-speaking African Americans. The proposed methodology is flexible enough to handle areal or geo-statistical datasets, and hierarchical models with multiple random intercepts.

email: lfboehm@gmail.com

SPATIAL-TEMPORAL MODELING OF THE CRITICAL WINDOWS OF AIR POLLUTION EXPOSURE FOR PRETERM BIRTH

Joshua Warren*, University of North Carolina, Chapel Hill
Montserrat Fuentes, North Carolina State University
Amy Herring, University of North Carolina, Chapel Hill
Peter Langlois, Texas Department of State Health Services

Exposure to high levels of air pollution during the pregnancy is associated with increased probability of preterm birth (PTB), a major cause of infant morbidity and mortality. New statistical methodology is required to specifically determine when a particular pollutant impacts the PTB outcome, to determine the role of different pollutants, and to characterize the spatial variability in these results. We introduce a new Bayesian spatial model for PTB which identifies susceptible windows throughout the pregnancy jointly for multiple pollutants while allowing these windows to vary continuously across space and time. A directional Bayesian approach is implemented to correctly characterize the uncertainty of the climatic and pollution variables throughout the modeling process. We apply our methods to geo-coded birth outcome data from the state of Texas (2002-2004). Our results indicate the susceptible window for higher preterm probabilities is mid-first trimester for the fine PM and beginning of the first trimester for the ozone.

email: joshuawa@email.unc.edu

HETEROSCEDASTIC VARIANCES IN AREALLY REFERENCED TEMPORAL PROCESSES WITH AN APPLICATION TO CALIFORNIA ASTHMA HOSPITALIZATION DATA

Harrison S. Quick*, University of Minnesota
Sudipto Banerjee, University of Minnesota
Bradley P. Carlin, University of Minnesota

Often in regionally aggregated spatial models, a single variance parameter is used to capture variability in the spatial association structure of the model. In real world phenomena, however, spatially-varying factors such as climate and geography may impact the variability in the underlying process. Here, our interest is in modeling monthly asthma hospitalization rates over an 18 year period in the counties of California. Earlier work has accounted for both spatial and temporal association using a process-based method that permits inference on the underlying temporal rates of change, or gradients, and has revealed progressively muted transitions into and out of the summer months. We extend this methodology to allow for region-specific variance components and separate, purely temporal processes, both of which we believe can simultaneously help avoid over- and undersmoothing in our overall spatiotemporal process and our temporal gradient process. After demonstrating the effectiveness of our new model via simulation, we reanalyze the asthma hospitalization data and compare our findings to those from previous work.

email: quic0038@umn.edu

BAYESIAN SEMIPARAMETRIC MODEL FOR SPATIAL INTERVAL-CENSORED DATA

Chun Pan*, University of South Carolina
Bo Cai, University of South Carolina
Lianming Wang, University of South Carolina
Xiaoyan Lin, University of South Carolina

Interval-censored survival data are often recorded in medical practice. Although some methods have been developed for analyzing such data, issues still remain in terms of efficiency and accuracy in estimation. In addition, interval-censored data with spatial correlation are not unusual but less studied. In this paper, we propose an efficient Bayesian approach under proportional hazards model to analyze general interval-censored survival data with spatial correlation. Specifically, a linear combination of monotone splines is used to model the unknown baseline cumulative hazard function, leading to a finite number of parameters to estimate while maintaining adequate modeling flexibility. A two-step data augmentation through Poisson latent variables is used to facilitate the computation of posterior distributions that are essential in the MCMC sampling algorithm proposed. A conditional autoregressive distribution is employed to model the spatial dependency. A simulation study is conducted to evaluate the performance of the proposed method. The approach is illustrated through a geographically referenced smoking cessation data in southeastern Minnesota where time to relapse is modeled and spatial structure is examined.

email: chunpan2003@hotmail.com

SPATIO-TEMPORAL WEIGHTED ADAPTIVE DECONVOLUTION MODEL TO ESTIMATE THE CEREBRAL BLOOD FLOW FUNCTION IN DYNAMIC SUSCEPTIBILITY CONTRAST MRI

Jiaping Wang*, University of North Texas, Denton
Hongtu Zhu, University of North Carolina, Chapel Hill
Hongyu An, University of North Carolina, Chapel Hill

Dynamic susceptibility contrast MRI measures the perfusion in numerical diagnostic and therapy-monitoring settings. One approach to estimate the blood flow parameters assume a convolution relation between the arterial input function and the tissue enhancement profile of the ROIs via a residue function, then extract the residue functions by some deconvolution techniques like the singular value decomposition (SVD), or Fourier transform based method for each voxel independently. This paper develops a spatio-temporal weighted adaptive deconvolution model (SWADM) to estimate these parameters by accounting for the complex spatio-temporal dependence and patterns in the images adaptively. SWADM has three features: being spatial, being hierarchical, being adaptive. To hierarchically and spatially denoise functional images, SWADM creates adaptive ellipsoids at each location to capture spatio-temporal dependence among imaging observations in neighboring voxels and times. A simulation study is used to demonstrate the meth and examine its finite sample performance. Our simulation study confirms that SWADM outperforms the voxel-wise deconvolution approach and SVD. Then the method is applied to the real data and compared with the results from the voxel-wise approach and SVD.

email: jwang@bios.unc.edu

47. INNOVATIVE DESIGN AND ANALYSIS ISSUES IN FETAL GROWTH STUDIES

CLINICAL IMPLICATIONS OF THE NATIONAL STANDARD FOR NORMAL FETAL GROWTH

S. Katherine Laughon*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Normal fetal growth is a marker of an optimal intrauterine environment and is important for the long-term health of the offspring. Defining normal and abnormal fetal growth in clinical practice and research has not been straightforward. Many clinical and epidemiologic studies have classified abnormal growth as small for gestational age (SGA) or large for gestational age (LGA) using normative birth weight references that may not reflect patterns of under- or overgrowth. An SGA neonate may be constitutionally small, while a normal birth weight percentile can occur in the setting of suboptimal fetal growth. In addition, only several longitudinal ultrasound studies have been conducted, and most of the larger studies were performed in Europe with the majority of subjects being Caucasian. I will discuss the study design for the NICHD Fetal Growth Study which is enrolling 2400 women to represent the diversity of the United States. I will also discuss the importance of developing biostatistical methodology for establishing both distance (cross-sectional) and velocity (longitudinal) reference curves. Further, I will discuss the methodological challenges in establishing personalized reference curves and predictions of abnormal birth outcomes. This talk will serve as a prelude to more technical talks on statistical model development and design.

email: laughonsk@mail.nih.gov

STATISTICAL MODELS FOR FETAL GROWTH

Robert W. Platt*, McGill University

Statistical models for longitudinal measures of fetal weight as a function of gestational age provide important information for the study of fetal and infant outcomes. I will first discuss existing models for the relation between fetal weight and gestational age, their strengths and limitations, and statistical criteria by which one may evaluate model performance. I will then outline a linear mixed model, with a Box-Cox transformation of fetal weight values, and restricted cubic splines, to flexibly but parsimoniously model median fetal weight. I will systematically compare our model to other proposed approaches. All proposed methods are shown to yield similar median estimates, as evidenced by overlapping pointwise confidence bands, except after 40 completed weeks, where our method seems to produce estimates more consistent with observed data and clinical expectations. I will then demonstrate some statistical properties of so-called customized fetal growth measures, and show that these methods provide relatively modest gains in prediction of key outcomes.

email: robert.platt@mcgill.ca

SOME ANALYTICAL CHALLENGES OF THE INTERGROWTH-21ST PROJECT IN THE DEVELOPMENT OF FETAL GROWTH REFERENCE STANDARDS

Eric O. Ohuma*, University of Oxford
Jose Villar, University of Oxford
Doug G. Altman, University of Oxford

The INTERGROWTH-21st Project's primary objective is the production of international standards to describe fetal growth, postnatal growth of term and pre-term infants and the relationship between birth weight, length, head circumference and gestational age. Eight geographically defined sites across the world (in Brazil, China, India, Oman, Kenya, UK, USA and Italy) are participating in three major studies. The main aim of the Fetal Growth Longitudinal Study (FGLS) is the development of prescriptive standards for fetal growth using longitudinal data from 4,500 fetuses in the 8 study sites. A reliable estimate of gestational age (GA) can be obtained from women with regular 28-32 day menstrual cycles who are certain of the first day of their last menstrual period (LMP); otherwise CRL is used for dating. The INTERGROWTH-21st Project includes 4,500 women whose LMP and CRL estimates of GA agreed within 7 days. We aim to develop a new centile chart for estimating GA from CRL. The main statistical challenge is modelling data where the outcome variable, GA, is truncated at both ends by design, i.e. at 9 and 14 weeks. Three approaches that attempt to overcome the truncation are evaluated in a simulation study and applied to our data.

email: eric.ohuma@obs-gyn.ox.ac.uk

LINEAR MIXED MODELS FOR REFERENCE CURVE ESTIMATION AND PREDICTION OF POOR PREGNANCY OUTCOMES FROM LONGITUDINAL ULTRASOUND DATA

Paul S. Albert*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
SungDuk Kim, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Linear mixed models can be used for reference curve estimation as well as predicting poor pregnancy outcomes from longitudinal ultrasound data. However, this class of models makes the strong assumption that individual heterogeneity is characterized by a Gaussian random effect distribution. Through simulations and asymptotic bias computations, we show situations where inference is insensitive to the Gaussian random effects assumption, while in other situations inferences rely heavily on this assumption. In situations where inferences are sensitive to the random effects assumptions, we show how alternative modeling approaches can be adapted to this problem. Further, we show how different models can be extended to allow for high-dimensional ultrasound and biomarker data for prediction and inference. Methodological results are illustrated with data from the Scandinavian Fetal Growth Project and the more recent NICHD National Fetal Growth Studies.

email: albertp@mail.nih.gov

48 BAYESIAN METHODS FOR MODELING MARK-RECAPTURE DATA WITH NON-INVASIVE MARKS

NON-INVASIVE GENETIC MARK-RECAPTURE WITH HETEROGENEITY AND MULTIPLE SAMPLING OCCASIONS

Janine A. Wright*, University of Otago, New Zealand
Richard J. Barker, University of Otago, New Zealand
Matthew R. Schofield, University of Kentucky

Wright et al. (2009) describes a method based on Bayesian imputation that allows modeling of data from non-invasive genetic samples in the presence of genotyping error in the form of allelic dropout. Wright et al model three processes that influence the observed data: the sampling process, genotype allocation to samples and the corruption of true genotypes through genotyping error. More flexibility in modeling the underlying biological situation is necessary and this can be catered for by modification of terms in the model corresponding to each of these processes. With respect to sampling, the assumption of equal probability of capture for all individuals can be overly simplistic and modification of the model allows for heterogeneity in capture probability. The model can be extended to allow for multiple sampling occasions and incorporating covariates (e.g. sample location) gives more information for resolving ambiguity in identification. Additionally, we can model open populations, with the aim of better parameter estimation and the ability to study relationships among individuals.

email: jwright@maths.otago.ac.nz

LATENT MULTINOMIAL MODELS

William A. Link*, U.S.Geological Survey Patuxent Wildlife Research Center

A variety of data types can be viewed as reflecting latent multinomial structure. For example, Yoshizaki et al. (2009) and Link et al. (2010) considered mark-recapture data with misidentifications. In their model M_{t,alpha} count frequencies are affine combinations of latent multinomial random variables. This talk presents methods for Bayesian analysis of such data, and useful tricks for estimating latent multinomial structure using standard software such as BUGS and JAGS.

email: wlink@usgs.gov

APPLICATION OF THE LATENT MULTINOMIAL MODEL TO DATA FROM MULTIPLE NON-INVASIVE MARKS

Simon J. Bonner*, University of Kentucky
Jason A. Holmberg, ECOCEAN USA

In some mark-recapture experiments, animals may be identified from multiple non-invasive marks. Combining the data from all marks has the potential to improve inference, but complications arise if the marks for a single individual cannot be matched without external information. Our work is motivated by data from the ECOCEAN worldwide study of whale sharks (Rhincodon typus), which uses skin patterns extracted from photographs submitted to an on-line database to identify individual sharks. These photographs may show either the left or right side of an individual, and the skin pigmentation patterns from the two sides of a shark cannot be matched unless photographs from both side are taken together on at least one encounter. Naively constructing capture histories from all photographs risks duplicating individuals (breaking the assumption of independence) but constructing capture histories from photographs of one side only reduces the amount of information in the data. This talk describes an extension of the latent multinomial model that properly accounts for data from multiple, natural marks. We present results from the whale shark analysis and a simulation study comparing the new model with the simple approaches that restrict data to a single mark or include all marks without adjusting for overcounting.

email: simon.bonner@uky.edu

49. HUNTING FOR SIGNIFICANCE IN HIGH-DIMENSIONAL DATA

DISCOVERING SIGNALS THROUGH NONPARAMETRIC BAYES

Linda Zhao*, University of Pennsylvania

Many classification problems can be conveniently formulated in terms of Bayesian mixture prior models. The mixture prior structure lends itself especially well for adapting to varying degrees of sparsity. Typically, parametric assumptions are made about the components of the mixture priors. In the following, we propose a parametric and a nonparametric classification procedures using a mixture prior Bayesian approach for a risk function that combines misclassification loss and an L_2 penalty. While the parametric procedure is closer to traditional approaches, in simulations, we show that the nonparametric classifier typically outperforms it when the parametric prior is misspecified; the two procedures have comparable performance even when the shape of the parametric prior is specified correctly. We illustrate the properties of the two classifiers on a publicly available gene expression dataset. This is a joint work with Fuki, I. and Raykar, V.

email: lzhao@wharton.upenn.edu

OPTIMAL MULTIPLE TESTING PROCEDURE FOR LINEAR REGRESSION MODEL

Jichun Xie, Temple University
Zhigen Zhao*, Temple University

Multiple testing problems are thoroughly understood for independent normal vectors but remain vague for dependent data. In this paper, we construct an optimal multiple testing procedure for the linear regression model when the dimension p is much larger than the sample size n. Linear regression model can be viewed as a generalization of the normal random vector model with an arbitrary dependency structure. The proposed procedure can control FDR under any specified levels; meanwhile, it can asymptotically minimize the FNR. In other word, it an achieve validity and efficiency at the same time. The numerical study shows that the proposed procedure has better performance compared with the competitive methods. We also applied the procedure to a genome-wide association study of hypertension for American African population and got interesting results.

email: zhaozhg@temple.edu

AN FDR APPROACH FOR MULTIPLE CHANGE-POINT DETECTION

Ning Hao, University of Arizona

The detection of change points has attracted a great deal of attention in many fields. From the hypothesis testing perspective, a multiple change-point problem can be viewed as a multiple testing problem by testing every data point as a potential change point. The false discovery rate (FDR) approach to multiple testing problems has been studied extensively since the seminal paper of Benjamini and Hochberg (1995). However, the multiple testing problem derived from change-point detection presents a problem that beyond the classical framework. In this talk, we will introduce an FDR approach for change-point detection, based on a screening and ranking algorithm. Both simulated and real data analyses will be presented to demonstrate the use of the SaRa.

ninghao008@gmail.com

ESTIMATION OF FDP WITH UNKNOWN COVARIANCE DEPENDENCE

Jianqing Fan; Princeton University
Xu Han*, Temple University

Multiple hypothesis testing is a fundamental problem in high dimensional inference, with wide applications in many scientific fields. When test statistics are correlated, false discovery control becomes very challenging under arbitrary dependence. In Fan, Han & Gu (2011), the authors gave a method to consistently estimate false discovery proportion when the covariance matrix of test statistics is known. The method is based on eigenvalues and eigenvectors of the covariance matrix. However, in practice the covariance matrix is usually unknown. Consistent estimate of an unknown covariance matrix is itself a difficult problem. In the current paper, we will derive some results to consistently estimate FDP even when the covariance matrix is unknown.

email: hanxu3@temple.edu

50. NEW DEVELOPMENTS IN THE CONSTRUCTION AND OPTIMIZATION OF DYNAMIC TREATMENT REGIMES

COVARIATE-ADJUSTED COMPARISON OF DYNAMIC TREATMENT REGIMES IN SEQUENTIALLY RANDOMIZED CLINICAL TRIALS

Xinyu Tang, University of Arkansas for Medical Sciences
Abdus S. Wahed*, University of Pittsburgh

Cox proportional hazards model is widely used in survival analysis to allow adjustment for baseline covariates. The proportional hazard assumption may not be valid for treatment regimes that depend on intermediate responses to prior treatments received, and it is not clear how such a model can be adapted to clinical trials employing more than one randomization. Besides, since treatment is modified post-baseline, the hazards are unlikely to be proportional across treatment regimes. Although Lokhnygina and Helterbrand (Biometrics. 2007 Jun;63(2):422-8.) introduced the Cox regression method for two-stage randomization designs, their method can only be applied to test the equality of two treatment regimes that share the same maintenance therapy. Moreover, their method does not allow auxiliary variables to be included in the model nor does it account for treatment effects that are not constant over time. In this article, we propose a model that assumes proportionality across covariates within each treatment regime but not across treatment regimes. Comparisons among treatment regimes are performed by testing the log ratio of the estimated cumulative hazards. The ratio of the cumulative hazard across treatment regimes is estimated using a weighted Breslow-type statistic. A simulation study was conducted to evaluate the performance of the estimators and proposed tests.

email: wahed@pitt.edu

ADAPTIVE TREATMENT POLICIES FOR INFUSION STUDIES

Brent A. Johnson*, Emory University

In post-operative medical care, some drugs are administered intravenously through an infusion pump. Comparing infusion drugs and rates of infusion are typically conducted through randomized controlled clinical trials across two or more arms and summarized through standard statistical analyses. However, the presence of infusion-terminating events can adversely affect primary endpoints and complicate statistical analyses of secondary endpoints. A secondary analysis of considerable interest is to assess the effects of infusion length once the test drug has been shown superior to standard of care. This analysis is complicated due to presence or absence of treatment-terminating events and potential time-varying confounding in treatment assignment. Connections to dynamic treatment regimes offer a principled approach to this secondary analysis and related problems, such as adaptive, personalized infusion policies, robust and efficient estimation, and the analysis of infusion policies in continuous time. These concepts will be illustrated with data from Duke University Medical Center.

email: bajohn3@emory.edu

NEAR OPTIMAL RANDOM REGIMES
James M. Robins*, Harvard School of Public Health

In high dimensional settings the search space for treatment strategies is too large to rely on optimal regime structural nested models fit by backward induction; yet ranking of candidate regimes by the estimated value functions is not possible because of the small probability of following a given regime. In a 2004 paper I discussed whether searching for near optimal random regimes has potential to help get one out of this quandary. I review recent further developments of this idea.

email: robins@hsph.harvard.edu

TARGETED LEARNING OF OPTIMAL DYNAMIC RULES
Mark J. van der Laan*, University of California, Berkeley

We present a targeted maximum likelihood estimator of the unknown parameters of a marginal structural working model for a class of dynamic or stochastic regimens. We present some data analysis results for rules for switching to another regimen for treating HIV infected patients. We also discuss an approach for targeted learning of the actual optimal rule among a given class of rules, using loss-based super learning.

email: laan@berkeley.edu

51. NOVEL BIOSTATISTICAL TOOLS FOR CURRENT PROBLEMS IN NEUROIMAGING

BAYESIAN SPATIAL VARIABLE SELECTION AND CLUSTERING FOR FUNCTIONAL MAGNETIC RESONANCE IMAGING DATA ANALYSIS
Fan Li, Duke University
Tingting Zhang*, University of Virginia

We develop a Bayesian variable selection framework for inferring the relationship between individual traits and brain activity from multi-subject fMRI data. The estimates of the brain hemodynamic responses for each voxel are used as predictors in a regression model where the response is the individual trait(s). This framework achieves identification and clustering of active brain regions simultaneously via a Dirichlet process prior with a spike and slab base measure, and also incorporates spatial information between brain voxels into the variable selection process via an Ising prior.

email: tz3b@virginia.edu

MODELING THE EVOLUTION OF BRAIN RESPONSE
Mark Fiecas*, University of California, San Diego
Hernando Ombao, University of California, Irvine

Statistical models for analyzing fMRI time courses must correctly account for the behavior of the amplitude of the hemodynamic response function (HRF) over the course of a multi-trial fMRI experiment in order to give valid inference about brain activity. Existing methods either assume that all trials yield identical realizations of the the data, or use parametric modulation, which assumes that the HRF behaves in a parametric manner, with the parametric form specified a priori. We propose a nonparametric method for estimating the evolution of the HRF over the course of the experiment. We estimate the amplitude of the HRF using nonparametric regression in order to model both the evolution of the HRF over the course of the experiment and the correlation between the trials. Unlike parametric modulation, our proposed method is completely data-driven, and so our model can better capture the evolution of the HRF and, consequently, yield a more valid inference about brain activity. We illustrate our proposed method using fMRI data collected from a learning experiment.

email: mfiecas@ucsd.edu

SPARSE AND FUNCTIONAL PCA WITH APPLICATIONS TO NEUROIMAGING
Genevera I. Allen*, Rice University and Baylor College of Medicine

Regularized principal components analysis, especially Sparse PCA and Functional PCA, has become widely used for dimension reduction in high-dimensional settings. Many examples of massive data, however, may benefit from estimating both sparse AND functional factors. These include neuroimaging data where there are discrete brain regions of activation (sparsity) but these regions tend to be smooth spatially (functional). Here, we introduce a framework for regularization of PCA that can encourage both sparsity and smoothness of the row and/or column PCA factors. This framework generalizes many of the existing optimization problems used for Sparse PCA, Functional PCA and two-way Sparse PCA and Functional PCA, as these are all special cases of our method. In particular, our method permits flexible combinations of sparsity and smoothness that lead to improvements in feature selection and signal recovery as well as more interpretable PCA factors. We demonstrate the utility of these methods on simulated data and a neuroimaging example on electroencephalography (EEG) data.

email: gallen@rice.edu

FUNCTIONAL DATA ANALYSIS FOR fMRI
Martin A. Lindquist*, Johns Hopkins University

In recent years there have been a number of exciting developments in the area of functional data analysis (FDA). Many of the methods that have been developed are ideally suited for the analysis of functional Magnetic Resonance Imaging (fMRI) data, which consists of images and/or curves. In this talk we discuss how methods from FDA can be used to uncover exciting new results, not readily apparent using standard analysis techniques, in wide ranging areas of fMRI research such as the estimation of the hemodynamic response function, brain connectivity, prediction and the analysis of multi-modal data.

email: mlindqui@jhsph.edu

52. DESIGNS AND INFERENCES FOR CAUSAL STUDIES

ESSENTIAL CONCEPTS FOR CAUSAL INFERENCE IN RANDOMIZED EXPERIMENTS AND OBSERVATIONAL STUDIES IN BIOSTATISTICS
Donald B. Rubin*, Harvard University

Tracing the evolution of randomized experiments and observational studies for causal effects begins with the delineation of the essential concepts from Fisher and Neyman in the early 20th century. A description of the historical and damaging domination of routine models for data analysis, such as least squares regression and its companions (e.g., logistic regression) follows. One of the negative consequences of this domination was the focus on the push-button analysis of existing data sets rather than the thoughtful design of data collection and assembly for causal inference. Of utmost importance, observational studies for causal inference should be designed to approximate well-conducted randomized trials, where a key role is played by the estimated propensity score. Once a study is well-designed, analyses estimating causal effects should typically focus on the stochastic (multiple) imputation of the missing potential outcomes using modern computational tools, thereby allowing relevant estimands to be defined and robustly estimated. In complicated situations that need principal stratification on intermediate outcomes to define the relevant estimands, using direct likelihood inference to assess competing models can be a most effective path for arriving at parsimonious causal inferences.

email: typetwoerror@gmail.com

ASSESSING THE EFFECT OF TRAINING PROGRAMS USING NONPARAMETRIC ESTIMATORS OF DOSE-RESPONSE FUNCTIONS: EVIDENCE FROM JOB CORPS DATA

Michela Bia*, CEPS/INSTEAD, Luxembourg
Alessandra Mattei, University of Florence, Italy
Carlos Flores, University of Miami
Alfonso Flores-Lagunes, Binghamton University, State University of New York

In this paper we propose three semiparametric estimators of the dose-response function based on kernel and spline techniques. In many observational studies treatment may not be binary or categorical. In such cases, one may be interested in estimating the dose-response function in a setting with a continuous treatment. This approach relies on the uncounfoundedness assumption, which requires the potential outcomes to be independent of the treatment conditional on a set of covariates. In this context the generalized propensity score can be used to estimate dose-response functions (DRF) and marginal treatment effect functions. We evaluate the performance of the proposed estimators using Monte Carlo simulation methods. We also apply our approach to the problem of evaluating job training program for disadvantaged youth in the United States (Job Corps program). In this regard, we provide new evidence on the intervention effectiveness by uncovering heterogeneities in the effects of Job Corps training along the different lengths of exposure.

email: michela.bia@ceps.lu

CASE DEFINITION AND DESIGN SENSITIVITY

Dylan Small*, University of Pennsylvania
Jing Cheng, University of California, San Francisco
Betz Halloran, University of Washington and Fred Hutchinson Cancer Research Center
Paul Rosenbaum, University of Pennsylvania

In case control studies, there may be several possible definitions of a case, some narrower and some broader. Because unmeasured confounding is an inevitable concern in case-control studies, we would like to choose a case definition that is as insensitive to bias from unmeasured confounding as possible. The design sensitivity of a case definition is the maximum amount of bias there can be for which we can distinguish a treatment effect without bias from the bias in large samples. We study the impact of the narrowness of the case definition on design sensitivity. We develop a formula for this design sensitivity, present simulation studies to assess how accurately design sensitivity reflects the finite sample properties of different case definitions and analyze an empirical example.

email: dsmall@wharton.upenn.edu

BAYESIAN INFERENCE FOR A NON-STANDARD REGRESSION DISCONTINUITY DESIGN WITH APPLICATION TO ITALIAN UNIVERSITY GRANTS

Fan Li*, Duke University
Alessandra Mattei, University of Florence, Italy
Fabrizia Mealli, University of Florence, Italy

Regression discontinuity (RD) designs identify causal effects of interventions by exploiting treatment assignment mechanisms that are discontinuous functions of observed covariates. In standard RD designs, the probability of treatment changes discontinuously if a covariate exceeds a threshold. Motivated by the evaluation of Italian university grants, this article considers a non-standard RD setup where the treatment assignment is determined by both a covariate and an application status. In particular, we focus on a fuzzy RD design with this setup, where the causal estimand and estimation strategies are different from those in the standard instrumental variable approach to fuzzy RDs. A Bayesian approach is developed to draw inferences for the causal effects at the threshold. Multivariate outcomes are utilized to further sharpen the analysis. We apply the method to evaluate the effects of Italian university grants on student dropout and academic performances. Posterior predictive model checks and sensitivity analysis are also conducted to validate the analysis.

email: fli@stat.duke.edu

53. RECENT ADVANCES IN THE ANALYSIS OF MEDICAL COST DATA

ESTIMATES AND PROJECTIONS OF THE COST OF CANCER CARE IN THE UNITED STATES

Angela Mariotto*, National Cancer Institute, National Institutes of Health
Robin Yabroff, National Cancer Institute, National Institutes of Health

The economic burden of cancer in the United States is substantial and expected to significantly increase in the future because of the expected growth and aging of the population and improvements in survival as well as trends in treatment patterns and costs of care following cancer diagnosis. We will present a method that uses the Surveillance Epidemiology and End Results (SEER) data linked to Medicare claims in order to estimate and project the total direct medical costs of cancer. This method combines estimates and projections of US cancer prevalence by phases of care with average annual medical costs estimated from claims data. The method allows for sensitivity analysis of assumptions of future trends in population, incidence, survival and costs. In all of the sensitivity analysis scenarios we assumed a dynamic population increase as projected by the US Bureau of the Census, and used the most recently available data to estimate incidence, survival, and cost of cancer care.

email: mariotta@mail.nih.gov

GENERALIZED REDISTRIBUTE-TO-THE-RIGHT ALGORITHM: APPLICATION TO THE ANALYSIS OF CENSORED COST DATA

Shuai Chen, Texas A&M University
Hongwei Zhao*, Texas A&M Health Science Center

Costs assessment and cost-effectiveness analysis serve as an essential part in economic evaluation of medical interventions. In clinical trials and many observational studies, costs as well as survival data are frequently censored. Standard techniques for survival-type data are often invalid in analyzing censored cost data, due to the induced dependent censoring problem (Lin et al., 1997). In this talk, we will first examine the equivalency between a redistribute-to-the right (RR) algorithm and the popular Kaplan-Meier method for estimating the survival function of time (Efron, 1967). Next, we will extend the RR algorithm to the problem of estimating mean costs with censored data, and propose a simple RR (Pfeifer and Bang, 2005) and an efficient RR estimator. We will establish the equivalency between the RR estimators and some existing cost estimators derived from the inverse-probability-weighting technique and semiparametric efficiency theory. Finally, we will extend the RR algorithm to the problem of estimating the survival function of health costs, and conduct simulation studies to compare a new RR survival estimator with some existing survival estimators for costs.

email: zhao@srph.tamhsc.edu

CENSORED COST REGRESSION MODELS WITH EMPIRICAL LIKELIHOOD

Gengsheng Qin*, Georgia State University
Xiao-hua Zhou, University of Washington
Huazhen Lin, Sichuan University
Gang Li, University of California, Los Angeles

In many studies of health economics, we are interested in the expected total cost over a certain period for a patient with given characteristics. Problems can arise if cost estimation models do not account for distributional aspects of costs. Two such problems are (1) the skewed nature of the data, and (2) censored observations. In this paper we propose an empirical likelihood (EL) method for constructing a confidence region for the vector of regression parameters, and a confidence interval for the expected total cost of a patient with the given covariates. We show that this new method has good theoretical properties and we compare its finite-sample properties with those of the existing method. Our simulation results demonstrate that the new EL-based method performs as well as the existing method when cost data are not so skewed, and outperforms the existing method when cost data are highly skewed. Finally, we illustrate the application of our method to a data set.

email: gqin@gsu.edu

SEMIPARAMETRIC REGRESSION FOR ESTIMATING MEDICAL COST TRAJECTORY WITH INFORMATIVE HOSPITALIZATION AND DEATH

Na Cai, Eli Lilly and Company
Wenbin Lu*, North Carolina State University
Hao Helen Zhang, University of Arizona
Jianwen Cai, University of North Carolina, Chapel Hill

We study longitudinal medical cost data that are subject to informative hospital visits and death. A new class of semiparametric regression models are proposed to estimate the entire medical cost trajectory by jointly modeling three processes: death time, hospital visit, and the cost. The underlying process dependence is characterized by latent variables, whose distributions are left completely unspecified and hence flexible to capture the complex association structure. We derive estimating equations for parameter estimation and inference. The resulting estimators are shown to be consistent and asymptotically normal. A resampling method is further developed for variance estimation. One common goal in medical cost data analysis is to get an accurate estimate of the lifetime medical cost. The proposed approach provides a convenient framework to estimate the cumulative medical cost up to any time point, which includes the lifetime cost as a special case. Simulation studies demonstrate promising performance of the new procedure, and one application to a chronic heart failure medical cost data from University of Virginia Health System illustrates its practical use.

email: lu@stat.ncsu.edu

54. RISK PREDICTION AND CLUSTERING OF GENETICS DATA

ENSEMBLE CLUSTERING WITH LOGIC RULES

Deniz Akdemir*, Cornell University

In this article, the logic rule ensembles approach to supervised learning is applied to the unsupervised or semi-supervised clustering. Logic rules which were obtained by combining simple conjunctive rules are used to partition the input space and an ensemble of these rules is used to define a similarity matrix. Similarity partitioning is used to partition the data in an hierarchical manner. We have used internal and external measures of cluster validity to evaluate the quality of clusterings or to identify the number of clusters.

email: da346@cornell.edu

HOW TO CLUSTER GENE EXPRESSION DYNAMICS IN RESPONSE TO ENVIRONMENTAL SIGNALS

Yaqun Wang*, The Pennsylvania State University
Meng Xu, Nanjing Forestry University
Zhong Wang, The Pennsylvania State University
Ming Tao, Brigham and Women's Hospital/Harvard Medical School
Junjia Zhu, The Pennsylvania State University
Li Wang, The Pennsylvania State University
Runze Li, The Pennsylvania State University
Scott A. Berceli, University of Florida
Rongling Wu, The Pennsylvania State University

Organisms usually cope with change in the environment by altering the dynamic trajectory of gene expression to adjust the complement of active proteins. The identification of particular sets of genes whose expression is adaptive in response to environmental changes helps to understand the mechanistic base of gene-environment interactions essential for organismic development. We describe a computational framework for clustering the dynamics of gene expression in distinct environments through Gaussian mixture fitting to the expression data measured at a set of discrete time points. We outline a number of quantitative testable hypotheses about the patterns of dynamic gene expression in changing environments and gene-environment interactions causing developmental differentiation. The future directions of gene clustering in terms of incorporation of the latest biological discoveries and statistical innovations are discussed. We provide a set of computational tools that are applicable to modeling and analysis of dynamic gene expression data measured in multiple environments.

email: yxw179@gmail.com

STATISTICAL METHODS FOR FUNCTIONAL METAGENOMIC ANALYSIS BASED ON NEXT GENERATION SEQUENCING DATA

Lingling An*, University of Arizona
Naruekamol Pookhao, University of Arizona
Hongmei Jiang, Northwestern University
Jiannong Xu, New Mexico State University

The advent of next-generation sequencing technologies has greatly promoted the field of metagenomics which studies genetic material recovered directly from an environment. However, due to the massive short DNA sequences produced by the new sequencing technologies, there is an urgent need to develop efficient statistical methods to rapidly analyze the massive sequencing data generated from microbial communities and to accurately detect the features/functions present in a metagenomic sample/community. In particular, there lack of statistical methods focusing on functional analysis of metagenoimics at the low level, i.e., more specific level. This study focuses on detecting all possible functional roles that are present in a metagenomic sample/community and at the low level. In this research we propose a statistical mixture model to describe the probability of short reads assigned to the candidate functional roles, with sequencing error taken into account. The proposed method is comprehensively tested in simulation studies. It is shown that the method is more accurate in assigning reads to relative functional roles, compared with other existing method in functional metagenomic analysis. The method is also employed to analyze two real data sets.

email: anling@email.arizona.edu

DETECTION FOR NON-ADDITIVE EFFECTS OF SNPs AT EXTREMES OF DISEASE-RISKS

Minsun Song*, National Cancer Institute, National Institutes of Health
Nilanjan Chatterjee, National Cancer Institute, National Institutes of Health

GWAS have led to discoveries of thousands of susceptibility SNPs. A crucial next step for post-GWAS is to characterize risk associated with multiple SNPs simultaneously. The starting point is often to assume SNPs affect disease risk in an additive fashion under some scale and test adequacy of such model based on goodness-of-fit (GOF) statistic. Several GOF statistics are available but face common limitation that they are not very sensitive at extremes of risk distributions where departure of joint risk from the underlying additive models may be actually more prominent in practice. We develop a new method for testing adequacy of an underlying assumed model for joint risk of a disease associated with multiple risk factors. To make test more sensitive for detecting departure near tails of risk distributions, we propose forming a test statistic based on squared Pearson residuals summed over only those individuals achieving certain risk threshold and then maximizing such test statistics over different risk-thresholds. We derive an asymptotic distribution for the proposed test statistic. Through simulations, we show the proposed procedure is much more powerful than popular Hosmer-Lemeshow and other GOF tests. As subjects with extreme risk may be impacted most from knowledge of their risk estimates, checking adequacy of risk models at extremes of risk is very important for clinical applications.

email: songm4@mail.nih.gov

PATHWAY SELECTION AND AGGREGATION USING MULTIPLE KERNEL LEARNING FOR RISK PREDICTION

Jennifer A. Sinnott*, Harvard University
Tianxi Cai, Harvard University

Attempts to predict risk using high dimensional genomic data can be made difficult by the large number of features and the potential complexity of the relationship between features and the outcome. Integrating prior biological knowledge into risk prediction with such data by grouping genomic features into pathways and networks reduces the dimensionality of the problem and could improve models by making them more biologically grounded and interpretable. Pathways could have complex signals, so our approach to model pathway effects should allow for this complexity. The kernel machine framework has been proposed to model pathway effects because it allows for nonlinear relationships within pathways; it has been used to make predictions for various types of outcomes from individual pathways. When multiple pathways are under consideration, we propose a multiple kernel learning approach to select important pathways and efficiently combine information across pathways. We derive our approach for a general survival modeling framework with a convex objective function, and illustrate its application under the Cox proportional hazards and accelerated failure time (AFT) models. Numerical studies with the AFT model demonstrate that this approach performs well in predicting risk. The methods are illustrated with an application to breast cancer data.

email: jsinnott@hsph.harvard.edu

ASSOCIATION ANALYSIS OF COMPLEX DISEASES USING TRIADS, PARENT-CHILD PAIRS AND SINGLETON CASES

Ruzong Fan*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Separate analysis of triad families or case-control data is routinely used in association study. Triad studies are important since they are robust in terms of less prone to false positive due to population structure. Case control design is widely used in association study and it is powerful but it is prone to false positive due to population structure. By doing separate analysis using triads or case-control data, it does not fully take the advantage of each study and it can be less powerful than a combined analysis. In this paper, we develop likelihood-based statistical models and likelihood ratio tests to test association between complex diseases and genetic markers by using combinations of full triads, parent-child pairs, and affected singleton cases for a unified analysis. By simulation studies, we show that the proposed models and tests are very robust in terms of type I error evaluations, and are powerful by empirical power evaluations. The methods are applied to analyze cleft palate data of TGFA gene of an Irish study.

email: fanr@mail.nih.gov

GENOTYPE CALLING AND HAPLOTYPING FOR FAMILY-BASED SEQUENCE DATA

Wei Chen*, University of Pittsburgh School of Medicine
Bingshan Li, Vanderbilt University Medical Center
Zhen Zeng, University of Pittsburgh School of Public Health
Serena Sanna, Centro Nazionale di Ricerca (CNR), Italy
Carlo Sidore, Centro Nazionale di Ricerca (CNR), Italy
Fabio Busonero, Centro Nazionale di Ricerca (CNR), Italy
Hyun Min Kang, University of Michigan
Yun Li, University of North Carolina, Chapel Hill
Gonçalo R. Abecasis, University of Michigan

Emerging sequencing technologies allow common and rare variants to be systematically assayed across the human genome in many individuals. In order to improve variant detection and genotype calling, raw sequence data are typically examined across many individuals. We describe a method for genotype calling in settings where sequence data are available for unrelated individuals and parent-offspring trios and show that modeling trio information can greatly increase the accuracy of inferred genotypes and haplotypes, especially on low to modest depth sequence data. Our method considers both linkage disequilibrium patterns and the constraints imposed by family structure when assigning individual genotypes and haplotypes. Using both simulations and real data, we show trios provide higher genotype calling and phasing accuracy across the frequency spectrum than the existing methods that ignores family structure. Our method can be extended to handle nuclear and multi-generational families in a computationally feasible manner. We anticipate our method will facilitate genotype calling and haplotype inference for many ongoing sequencing projects.

email: chenw8@hotmail.com

55. AGREEMENT MEASURES FOR LONGITUDINAL/SURVIVAL DATA

MUTUAL INFORMATION KERNEL LOGISTIC MODELS WITH APPLICATION IN HIV VACCINE STUDIES

Saheli Datta, Fred Hutchinson Cancer Research Center
Youyi Fong*, Fred Hutchinson Cancer Research Center
Georgia Tomaras, Duke University

We propose a mutual information kernel logistic model to study the effect of protein sequences. A mutual information kernel measures the similarity between two observations using a probability model. We use the profile hidden Markov model to model a protein sequence. A kernel logistic model models the effect of protein sequences as a random effect whose covariance matrix is parameterized by the kernel. To test the null hypothesis that the protein sequence has an effect on the outcome, we approximate the score test statistics with a chi-squared distribution and take the maximum over a grid of a scale parameter which only exists under the alternative hypothesis. A parametric bootstrap approach is used to obtain the reference distribution. We apply our method to the HIV-1 vaccine study to identify regions of the gp120 protein sequence where IgA antibody binding correlates with infection risk.

email: youyifong@gmail.com

ESTIMATION AND INFERENCE OF THE THREE-LEVEL INTRACLASS CORRELATION COEFFICIENT

Mat D. Davis*, University of Pennsylvania and Theorem Clinical Research
J. Richard Landis, University of Pennsylvania
Warren Bilker, University of Pennsylvania

Since the early 1900's, the intraclass correlation coefficient (ICC) has been used to quantify the level of agreement among different assessments on the same object. By comparing the level of variability that exists within subjects to the overall error, a measure of the agreement among the different assessments can be calculated. Historically, this has been performed using subject as the only random effect. However, there are many cases where other nested effects, such as site, should be controlled for when calculating the ICC to determine the chance corrected agreement adjusted for other nested factors. We will present a unified framework to estimate both the two-level and three-level ICC for continuous and categorical data. In addition, the corresponding standard errors and confidence intervals for both continuous and categorical ICC measurements will be presented. Finally, an example of the effect that controlling for site can have on ICC measures will be presented for subjects within genotyping plates comparing genetically determined race to patient reported race.

email: davismat@mail.med.upenn.edu

EFFECTS AND DETECTION OF RANDOM-EFFECTS MODEL MISSPECIFICATION IN GLMM

Shun Yu*, University of South Carolina, Columbia
Xianzheng (Shan) Huang, University of South Carolina, Columbia

We develop a diagnostic method for identifying the skewness of the true distribution of random effects in generalized linear mixed models with binary responses. We investigate large-sample properties of maximum likelihood estimators under different ways of misspecify-

ing the random-effect distribution based on different data structures. Finite-sample studies are conducted to explore operation characteristics of the proposed diagnostic test. Besides detecting the skewness of random effects, the test can also identify other types of misspecification. We compare the proposed test with the test in Tchetgen and Coull (2006). Finally, these tests are applied to data from a longitudinal respiratory infection study.

email: yu34@email.sc.edu

A DISCRETE SURVIVAL MODEL WITH RANDOM EFFECT FOR DESIGNING AND ANALYZING REPEATED LOW-DOSE CHALLENGE

Chaeryon Kang*, Fred Hutchinson Cancer Research Center
Ying Huang, Fred Hutchinson Cancer Research Center

Repeated low-dose challenge (RLD) design is important in HIV vaccine/prevention research. Compared to the challenge study with a single high dose, the RLD more realistically reflects the low-probability of HIV infection in human and provides more statistical power for detecting vaccine effects. Current methods for RLD design relies heavily on an assumption of homogeneous probability of infection among animals which, upon violation, can lead to invalid inference and underpowered study design. In the present study, we propose to fit a discrete survival model with random effect that allows for heterogeneity in the infection risk among animals and allows for variation of challenge doses in the study. Based on this model, we derived likelihood ratio test and estimators for vaccine's efficacy. Simulation study demonstrates good finite sample properties of the proposed method and its superior performance compared to existing methods. We evaluate statical power by varying sample sizes, the maximum number of challenges per animal, and strength of within-animal dependency through intensive simulation studies. Application of the proposed approach to a RLD challenge study in HIV vaccine trial for Rhesus Macaque shows a significant within-animal dependency in the probability of infection. The results of our study provide useful guidelines for future RLD experimental design.

email: ckang2@fhcrc.org

COVARIATE ADJUSTMENT IN ESTIMATING THE AREA UNDER ROC CURVE WITH PARTIALLY MISSING GOLD STANDARD

Danping Liu*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Xiao-Hua Zhou, University of Washington

In ROC analysis, covariate adjustment is advocated when the covariates impact the magnitude or accuracy of the test under study. For many large-scale screening tests, the true condition status may be missing because it is expensive and/or invasive to ascertain the disease status. The complete-case analysis may end up with a biased inference, also known as verification bias. To address the issue of covariate adjustment with verification bias in ROC analysis, we propose several estimators for the area under the covariate-specific and covariate-adjusted ROC curves (AUCx and AAUC). The AUCx is directly modeled in the form of binary regression, and the estimating equations are based on the U statistics. The AAUC is estimated from the weighted average of AUCx over the covariate distribution of the diseased subjects. We employ reweighting and imputation techniques to overcome the verification bias problem. Our proposed estimators are initially derived under the assumption of missing at random, and then with some modification, the estimators can be extended to the not-missing-at-random situation. The asymptotic distributions are derived for the proposed estimators. The finite sample performance is evaluated by a series of simulation studies. Our method is applied to a data set in Alzheimer's disease research.

email: danping.liu@nih.gov

NOVEL AGREEMENT MEASURES FOR CONTINUOUS SURVIVAL TIMES

Tian Dai*, Emory University
Ying Guo, Emory University
Limin Peng, Emory University
Amita K. Manatunga, Emory University

Assessment of agreement is often of interest in biological sciences when an outcome is measured on the same subject by different methods/raters. Most of the existing agreement measures are only suitable for complete observations and measure the global agreement. In survival studies, outcome of interest are usually subject to censoring and are often only observed in a limited region due to the restriction of the follow-up period. To address these issues, we propose new agreement measures for correlated survival times. We first develop a local agreement measure defined based on bivariate hazard functions which reflects the agreement between the survival outcomes at a specific time point. We then propose a regional agreement measure by summarizing the local measure over a finite region. The proposed measures can readily accommodate censored observations, reflect the pattern of agreement on the two-dimensional time plane, and also allow us to measure the agreement over a finite region of interest within the survival time space. Simulation studies and application to a survival data example would be discussed.

email: tian.dai88@gmail.com

56. IMAGING

GENETIC DISSECTION OF NEUROIMAGING PHENOTYPES

Yijuan Hu*, Emory University
Jian Kang, Emory University

There is a major interest in investigation of such diseases based on genetic and neuroimaging biomarkers. Recent advances in neuroimaging and genetics allow imaging genetics studies to collect both highly detailed brain images (>100,000 voxels) and genome-wide genotype information (>12 million known variants). However, there is lack of statistical powerful methods and computationally efficient tools to analyze such very high-dimensional data, which has greatly hindered the impact of imaging genetics studies. In this paper, we propose a statistical method to assess the association between genetic variants and neuroimaging traits. Our method is tailored to the brain-wide, genome-wide association discovery while accounting for the spatial correlation of imaging data and linkage disequilibrum of genetics markers. Simulation studies and the analysis of Alzheimer's Disease Neuroimaging Initiative (ADNI) data demonstrate that our method is computationally efficient and more powerful in detection of true genetic effects.

email: yijuan.hu@emory.edu

FITTING THE CORPUS CALLOSUM USING PRINCIPAL SURFACES

Chen Yue*, Johns Hopkins University
Brian S. Caffo, Johns Hopkins University
Vadim Zipunnikov, Johns Hopkins University
Dzung L. Pham, Radiology and Imaging Sciences, National Institutes of Health
Daniel S. Reich, National Institute of Neurological Disorders and Stroke, National Institutes of Health

In this manuscript we are concerned with a group of data generated from a diffusion tensor imaging (DTI) experiment. The goal is to evaluate corpus callosum properties and to relate these properties to multiple sclerosis disease status and progression. We approach the problem by finding a geometrically motivated surface-based representation of the corpus callosum and visualize the fractional anisotropy (FA) value projected onto the surface. Thus we describe an algorithm to construct the principal surface of a corpus callosum and project associated FA values onto the flattened surface. The algorithm has been tested on a variety of simulation cases. In our application, after finding the surface, we subsequently flattened it to obtain two dimensional visualizations of corpus callosum dif-

fusion properties. The algorithm has been implemented on 176 multiple sclerosis (MS) subjects observed at 466 visits (scans). For each subject and visit the study contains a registered DTI scan of the corpus callosum at roughly 20,000 voxels.

email: cyue@jhsph.edu

FAST SCALAR-ON-IMAGE REGRESSION WITH APPLICATION TO ASSOCIATION BETWEEN DTI AND COGNITIVE OUTCOMES

Lei Huang*, Johns Hopkins Bloomberg School of Public Health
Jeff Goldsmith, Columbia University Mailman School of Public Health
Philip T. Reiss, New York University School of Medicine
Daniel Reich, Johns Hopkins School of Medicine
Ciprian M. Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Tractography is a technique for estimating and quantifying brain pathways in vivo using diffusion tensor imaging (DTI) data. These pathways often subserve specific functions, and damage to them may result in characteristic forms of disability. Several standard regression techniques are reviewed here to quantify the relationship between such damage and the extent of disability. Traditional approaches focus on voxel-wise regressions that does not take into account the complex within-brain spatial correlations, and may fail to correctly estimate the form and size of the association. We propose a 'scalar-on-image' regression procedure to address these issues. Our procedure introduces a latent binary map that estimates the locations of predictive voxels corrected for the association between all other voxels and the outcome. By inducing a spatial prior structure the procedure produces a sparse association map which also maintains spatial continuity of predictive regions. The method is demonstrated on a large study of association between fractional anisotropy in a cross-sectional population of MRIs for 135 multiple-sclerosis patients and cognitive disability measures.

email: huangracer@gmail.com

INVESTIGATION OF STRUCTURAL CONNECTIVITY UNDERLYING FUNCTIONAL CONNECTIVITY USING fMRI DATA

Phebe B. Kemmer*, Emory University
Ying Guo, Emory University
F. DuBois Bowman, Emory University

A better understanding of brain functional connectivity (FC), structural connectivity (SC), and the relationship between these measures can provide important insights on neural representations underlying healthy and diseased brains. In this work, we aim to investigate the strength of SC underlying FC networks estimated from data-driven methods such as independent component analysis (ICA). We are also interested in examining whether the strength of SC is associated with the reliability of the FC networks. To achieve our goals, we propose a new statistical measure of the strength of Structural Connectivity (sSC) based on diffusion tensor imaging (DTI) data. The sSC measure provides a summary of SC strength within an identified FC network, while controlling for the network's baseline level of anatomical connectivity. As a standardized index, the sSC measure can be compared across FC networks of different sizes. The estimation method for the sSC measure would be discussed. We will illustrate the application of sSC using functional magnetic resonance imaging (fMRI) data.

email: pbrenne@emory.edu

NETWORK ANALYSIS OF RESTING-STATE fMRI USING PENALIZED REGRESSION MODELS

Gina D'Angelo*, Washington University School of Medicine
Gongfu Zhou, Washington University School of Medicine

Network analysis is an area of interest of resting-state fMRI. One of the objectives in this area is to identify various networks that exist to discriminate healthy populations from those with dementia. We propose using elastic net and least absolute shrinkage and selection operator (lasso) to find networks that differ across groups. The inter-regional correlations placed into the regression models will be estimated using a two-stage generalized estimating equations approach. We will also discuss various multiple comparison corrections and resampling approaches for inference. The method will be demonstrated using an Alzheimer's disease functional connectivity study. We will also perform simulation studies to study the properties of this approach.

email: gina@wubios.wustl.edu

EFFECTIVE CONNECTIVITY MODELING OF FUNCTIONAL MRI USING DYNAMIC CAUSAL MODELING WITH APPLICATION TO DEPRESSION IN ADOLESCENCE

Donald R. Musgrove*, University of Minnesota
Lynn E. Eberly, University of Minnesota
Kathryn R. Cullen, University of Minnesota

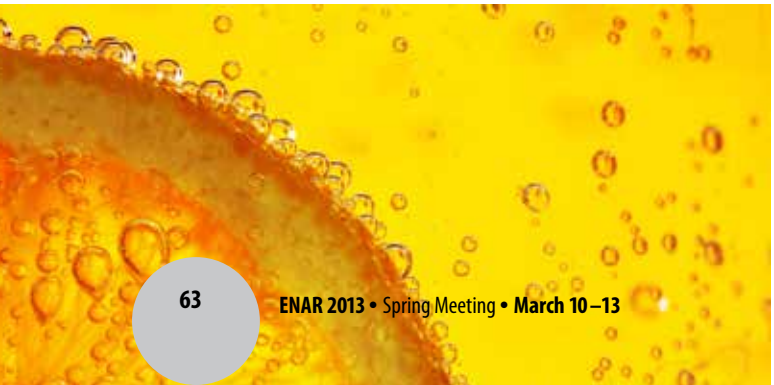
This project aims to examine the neural underpinnings of depression in adolescence. Researchers collected multimodal neuroimaging, including functional MRI with a task paradigm, in adolescents aged 12-19 years with and without depression. Effective connectivity analysis is used to estimate parameters that represent neural influences among regions of the brain with respect to the experimental tasks over time. The brain is modeled as a deterministic system whose inputs are the experimental manipulations with outputs as hemodynamic signals measured by fMRI. Dynamic causal modeling (DCM) is a bi-linear state space model for effective connectivity analysis that allows for evaluation of competing hypotheses of connectivity within the neural system. Unobserved changes in neuronal states over time are linked to observed fMRI data via a hemodynamic model; parameters driving the neural state changes are estimated using a Variational Bayes EM scheme. Effective connectivity using DCM requires per-participant evaluation and comparison of models representing competing hypotheses on the connective architecture of the investigated neural system. Group level comparisons are then carried out to investigate whether adolescents with/without depression have different architecture or different strengths of connections within the same architecture.

email: musgr007@umn.edu

LAPLACE DECONVOLUTION AND ITS APPLICATION TO DYNAMIC CONTRAST ENHANCED IMAGING

Marianna Pensky*, University of Central Florida
Fabienne Comte, University of Paris V
Yves Rozenholc, University of Paris V
Charles-Andre Cuenod, University of Paris V and European Hospital George Pompidou

In the present paper we consider the problem of Laplace deconvolution with noisy discrete observations. The study is motivated by Dynamic Contrast Enhanced imaging using a bolus of contrast agent, a procedure which allows considerable improvement in{evaluating} the quality of a vascular network and its permeability and is widely used in medical assessment of brain flows or cancerous tumors. Although the study is motivated by medical imaging application, we obtain a solution of a general problem of Laplace deconvolution based on noisy data which appears in many different contexts. We propose a new method



for Laplace deconvolution which is based on expansions of the convolution kernel, the unknown function and the observed signal over Laguerre functions basis. The advantage of this methodology is that it leads to very fast computations, does not require exact knowledge of the kernel and produces no boundary effects due to extension at zero and cut-off at T.

email: marianna.pensky@ucf.edu

57. STATISTICAL CONSULTING AND SURVEY RESEARCH

ANALYSIS OF POPULATION-BASED CASE-CONTROL STUDIES WITH COMPLEX SAMPLES ON HAPLOTYPE EFFECTS EXPLOITING GENE-ENVIRONMENT INDEPENDENCE IN GENETICS ASSOCIATION STUDIES

Daoying Lin*, University of Texas, Arlington
Yan Li, University of Maryland

The use of complex sampling in population-based case-control studies (PBCCS) is becoming more common, particularly for selection of controls that range from random digit dialing sampling in telephone surveys to stratified multistage sampling in household surveys and surveys of patients from physician practices. It is important to account for design complications in the statistical analysis of the PBCCS with complex sampling. Attracted by the efficiency advantage of the retrospective method, we explore the assumptions of Hardy-Weinberg Equilibrium (HWE) and gene-environment (G-E) independence in the underlying population. For the two-step approach, we also describe a variance estimation formula that could incorporate the uncertainty in haplotype frequency estimates which are ignored elsewhere. Results of our simulation studies demonstrate superior performance of the proposed methods under complex sampling over existing methods. An application of the proposed methods is illustrated using a population-based case-control study of kidney cancer. An R package has also been developed to conduct the relative analysis.

email: daoying.lin@mavs.uta.edu

CONDITIONAL PSEUDOLIKELIHOOD AND GENERALIZED LINEAR MIXED MODEL METHODS FOR TO ADJUST FOR CONFOUNDING DUE TO CLUSTER WITH ORDINAL, MULTINOMIAL, OR NONNEGATIVE OUTCOMES AND COMPLEX SURVEY DATA

Babette A. Brumback*, University of Florida
Zhuangyu Cai, University of Florida
Zhulin He, National Institute of Statistical Sciences and American Institutes for Research
Hao Zheng, University of Florida
Amy B. Dailey, Gettysburg College

In order to adjust individual-level covariate effects for confounding due to unmeasured neighborhood characteristics, we have recently developed conditional pseudolikelihood and generalized linear mixed model methods for use with complex survey data. The methods require sampling design joint probabilities for each within-neighborhood pair. For the conditional pseudolikelihood methods, the estimators and asymptotic sampling distributions we present can be conveniently computed using standard logistic regression software for complex survey data, such as SAS PROC SURVEYLOGISTIC. For the generalized linear mixed model methods, computation is straightforward using Stata's GLLAMM macro. We demonstrate validity of the methods theoretically, and also empirically using simulations. We apply the methods to data from the 2008 Florida Behavioral Risk Factor Surveillance System survey, in order to investigate disparities in frequency of dental cleaning both unadjusted and adjusted for confounding by neighborhood.

email: brumback@ufl.edu

LASSO-BASED METHODOLOGY FOR QUESTIONNAIRE DESIGN APPLIED TO THE OPPERA STUDY

Erika Helgeson*, University of North Carolina, Chapel Hill
Gary Slade, University of North Carolina, Chapel Hill
Richard Ohrbach; University of Buffalo
Roger Fillingim, University of Florida
Joel Greenspan, University of Maryland, Baltimore
Ron Dubner, University of Maryland, Baltimore
William Maixner, University of North Carolina, Chapel Hill
Eric Bair, University of North Carolina, Chapel Hill

In survey research, one may wish to predict a clinical outcome based on responses to the survey. One may wish to keep the questionnaire length reasonable without sacrificing predictive accuracy. In this study, we analyze data collected in the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study regarding comorbid pain conditions associated with Temporomandibular Disorders (TMD). A survey administered in OPPERA contains a 21-item checklist of comorbid pain conditions. Previous research has shown that a simple count of the number of comorbid pain conditions is associated with both chronic TMD and first-onset TMD. However, this scoring method requires the completion of a lengthy questionnaire. Our objective was to find a shorter version of the questionnaire without decreasing the ability to predict either of the primary outcomes of interest. We developed a shortened version of the questionnaire by fitting a lasso regression model to predict both chronic and first-onset TMD with indicators for each of the 21 questionnaire items as predictors. A subset of the 21-items was determined to be as strongly associated with both chronic and first-onset TMD as the full 21-item checklist. These results indicate that this method may be useful in other survey research studies.

email: helgeson@live.unc.edu

VARIABLE SELECTION AND ESTIMATION FOR LONGITUDINAL SURVEY DATA

Lily Wang*, University of Georgia
Suojin Wang, Texas A&M University

There is wide interest in studying longitudinal surveys where sample subjects are observed successively over time. Longitudinal surveys have been used in many areas today, for example, in the health and social sciences, to explore relationships or to identify significant variables in regression settings. We develop a general strategy for the model selection problem in longitudinal sample surveys. A survey weighted penalized estimating equation approach is proposed to select significant variables and estimate the coefficients simultaneously. The proposed estimators are design consistent and perform as well as the oracle procedure when the correct sub-model were known. The estimating function bootstrap is applied to obtain the standard errors of the estimated parameters with good accuracy. A fast and efficient variable selection algorithm is developed to identify significant variables for complex longitudinal survey data. Simulated examples are illustrated to show the usefulness of the proposed methodology under various model settings and sampling designs.

e-mail: lilywang@uga.edu

RATING SCALES AS PREDICTORS—THE OLD QUESTION OF SCALE LEVEL AND SOME ANSWERS

Jan Gertheiss*, Ludwig Maximilian University, Munich
Gerhard Tutz, Ludwig Maximilian University, Munich

Rating scales as predictors in regression models are typically treated as metrically scaled variables or, alternatively, are coded in dummy variables. The first approach implies a scale level that is not justified, the latter approach results in a large number of parameters to be estimated. Therefore, when dummy variables are used most applications are restricted to settings with only a few predictors. The penalization approach advocated here takes the scale level serious by using only the ordering of categories, but is shown to work in the high-dimensional

case. Moreover, our approach is also useful in lower dimensions, when association between a categorical predictor with ordered levels and a metric response variable is to be tested, or for testing whether the regression function is linear in the ordinal predictor's class labels.

e-mail: jan.gertheiss@stat.uni-muenchen.de

PRACTICAL ISSUES IN THE DESIGN AND ANALYSIS OF DUAL-FRAME TELEPHONE SURVEYS

Bo Lu*, The Ohio State University
Juan Peng, The Ohio State University
Timothy Sahr, The Ohio State University

With the setup of a large complex dual-framed telephone survey, we investigate the impact of different design and analytical strategies of dual-frame surveys for the estimation of population quantities. We compare estimation strategies via an extensive simulation study. For the simulation, two design options are considered, cell-only and cell-any, and several main estimation techniques are considered, simple composite estimation, single frame estimation, and pseudo maximum likelihood estimation. Both fixed size and fixed budget scenarios are examined in the simulation to take into account the differential cost of sampling in landline and cell phone populations. Results indicate that cell-only design is less biased than the cell-any design with no raking or poor raking. If accurate raking information is available for different phone use patterns, all estimation techniques provide similar results and the cell-any supplemental sampling is more cost efficient. We apply different estimation techniques to the Ohio Family Health Survey.

e-mail: blu@cph.osu.edu

EXPERIENCE WITH STATISTICAL CONSULTING ON GRANT SUBMISSION IN A LARGE MEDICAL CENTER

James D. Myles*, University of Michigan
Robert A. Parker, University of Michigan

Graduate programs in biostatistics focus on statistical methods rather than consulting. Since 2007, UM has provided free research development services to investigators before grant submission. In the last year 277 investigators requested services ranging from a letter of support, grant editing, help with budgeting and submission, and over a 100 full reviews. Full reviews consist of review written material and a face to face meeting with the investigator(s) by a senior clinician(s) and a statistician. Statistical contributions include suggestions on changing the aims of the study, changing the basic study design, referrals to other investigators, and occasional power calculations. It is common for the group to advise an investigator to abandon an application. Over 55 years ago Cochrane and Cox wrote that a statistician's major contribution to a study rarely involves subtle statistical issues. Far more impact is made by questioning investigators about why they are doing the experiment, how it will impact their field and how a completed experiment will be able to answer the question posed. Graduate training programs need to provide more real-world consulting experience to students.

e-mail: jdmyles@umich.edu

58. CATEGORICAL DATA METHODS

AN EFFICIENT AND EXACT APPROACH FOR DETECTING TRENDS WITH BINARY ENDPOINTS

Guogen Shan*, University of Nevada, Las Vegas

Lloyd (Aust. Nz. J. Stat., 50, 329 345, 2008) developed an exact testing approach to control for nuisance parameters, which was shown to be advantageous in testing for differences between two population proportions. We utilized this approach to obtain unconditional tests for trends in 2×K contingency tables. We compare the unconditional procedure with other unconditional and conditional approaches based on the well-known Cochran-Armitage test statistic. We give an example to illustrate the approach, and provide a comparison between the methods with regards to type I error and power. The proposed procedure is preferable because it is less conservative and has superior power properties.

email: guogen.shan@unlv.edu

PERMUTATION TESTS FOR SUBGROUP ANALYSES WITH BINARY RESPONSE

Siyoen Kil*, The Ohio State University
Eloise Kaizar, The Ohio State University

The goal of subgroup analyses in clinical trial is to quantify the heterogeneity of treatment effects across subpopulations. Identifying a subpopulation that benefits from treatment or whose benefit is enhanced over the population at large can improve healthcare for patients and targeted marketing for drug developers. Often times for subgroup analyses, evaluating the joint distribution among test statistics for multiple subgroups is difficult because the correlation is not known or not established by the multiple null hypotheses of no heterogeneity. Permutation testing is very tempting because one might expect the correlation structure to be retained by the permutation procedure while still controlling Type-I error at the nominal level. However, permuting subgroup membership labels does not produce a valid reference distribution for a randomized clinical trial with discrete outcome, even with only one subgroup classifier. We present some numerical studies that verify that permutation based reference distributions do not control type I error rate for common test statistics. These theoretical and numerical results also provide intuition for the null hypothesis corresponding to permutation testing in the subgroup analysis context.

email: kil.3@osu.edu

ARE YOU LOOKING FOR THE RIGHT INTERACTIONS? ADDITIVE VERSUS MULTIPLICATIVE INTERACTIONS WITH DICHOTOMOUS OUTCOME VARIABLES

Melanie M. Wall*, Columbia University
Sharon Schwartz, Columbia University

It is common in the health sciences to assess interactions (or effect measure modifications). The most common way to operationalize the estimation of this moderation effect is through the inclusion of a cross-product term between the exposure and the potential moderator in a regression analysis. This cross-product term is commonly called an interaction term. In the present talk we will emphasize that the answer to the question of whether there is or is not effect measure modification depends on what scale the interaction is considered. Specifically in the case of dichotomous outcomes this means whether we are examining risk differences (additive scale) or odds ratios or risk ratios (multiplicative scale). While it is common to test the statistical significance of the interaction term in a logistic regression, this is only a test for interaction on the logit (i.e. odds ratio - multiplicative) scale, and in general it is not consistent with the test for interaction on the probability (i.e. risk difference - additive) scale. Different methods with be compared for testing the interaction on the additive scale including 1) linear binomial regression 2) weighted least squares 3) logistic regression with marginal and conditional mean back-transformation. Worked examples will be shown.

email: mmwall@columbia.edu

COMPARISON OF ADDITIVE AND MULTIPLICATIVE BAYESIAN MODELS FOR LONGITUDINAL COUNT DATA WITH OVERDISPERSION PARAMETERS: A SIMULATION STUDY

Mehreteab F. Aregay*, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium
Ziv Shkedy, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium
Geert Molenberghs, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium

In applied statistical data analysis, overdispersion is a common feature. It can be addressed using both multiplicative and additive andom effects. A multiplicative model for count data enters a gamma random effect as a multiplicative factor into the mean, whereas an additive model assume a normally distributed random effect, entered into the linear predictor. Using Bayesian principles, these ideas are applied to longitudinal count data, based on the work of Molenberghs, Verbeke, and Dem´etrio (2007). The performance of the additive and multiplicative approaches is compared using a simulation study.

email: mehreteabfantahun.aregay@med.kuleuven.be

THE FOCUSED AND MODEL AVERAGE ESTIMATION FOR PANEL COUNT DATA

HaiYing Wang*, University of Missouri
Jianguo Sun, University of Missouri
Nancy Flournoy, University of Missouri

One of the main goals in model selection is to improve the quality of the estimators of interested parameters. For this, Claeskens and Hjort proposed the focused information criterion (FIC) which emphasis on the accuracy of estimation of particular parameters of interest. In a companion paper, they showed that the estimation efficiency can be further improved by taking a weighted average on sub-model estimators. The purpose of this paper is to extend the aforementioned ideas to panel count data. Panel count data frequently occurs in long-term medical follow-up studies, in which the primary object is often to evaluate the effectiveness of newly developed medicine or treatments. In terms of statistical modeling, the effectiveness is often depicted by only a few parameters, although the inclusion of other parameters and covariates affects the estimation of parameters of interest. So the focused and model average estimation fill the need of this problem ideally. In the context of panel count data, we define the FIC and derive the asymptotic distribution of the model average estimator. A simulation study is carried out to examine the finite sample performance and a real data from a cancer study is analyzed to illustrate the practical application.

email: hwzq7@mail.missouri.edu

TWO-SAMPLE NONPARAMETRIC COMPARISON FOR PANEL COUNT DATA WITH UNEQUAL OBSERVATION PROCESSES

Yang Li*, University of Missouri, Columbia
Hui Zhao, Huazhong Normal University, China
Jianguo Sun, University of Missouri, Columbia

This article considers two-sample nonparametric comparison based on panel count data. Most approaches that have been developed in the literature require an equal observation process for all subjects. However, such assumption may not hold in reality. A new class of test procedures are proposed that allow unequal observation processes for the subjects from different treatment groups, and both univariate and multivariate panel count data are considered. The asymptotic normality of the proposed test statistics is established and a simulation study is conducted to evaluate the finite sample properties of the proposed approach. The simulation results show that the proposed procedures work well for practical situations and especially for sparsely distributed data. They are applied to a set of panel count data from a skin cancer study.

email: ylx33@mail.missouri.edu

A MARGINALIZED ZERO-INFLATED POISSON REGRESSION MODEL WITH OVERALL EXPOSURE EFFECTS

D. Leann Long*, University of North Carolina, Chapel Hill
John S. Preisser, University of North Carolina, Chapel Hill
Amy H. Herring, University of North Carolina, Chapel Hill

The zero-inflated Poisson (ZIP) regression model is often employed in public health research to examine the relationships between exposures of interest and a count outcome exhibiting many zeros, in excess of the amount expected under Poisson sampling. The regression coefficients of the ZIP model have latent class interpretations that are not well suited for inference targeted at overall exposure effects, specifically, in quantifying the effect of an explanatory variable in the overall mixture population. We develop a marginalized ZIP model approach for independent responses to model the population mean count directly, allowing straightforward inference for overall exposure effects and easy accommodation of offsets representing individuals' risk times and empirical robust variance estimation for overall log incidence density ratios. Through simulation studies, the performance of maximum likelihood estimation of the marginalized ZIP model is assessed and compared to existing post-hoc methods for the estimation of overall effects in the traditional ZIP model framework. The marginalized ZIP model is applied to a recent study of a safer sex counseling intervention.

email: dllong@email.unc.edu

59. GRADUATE STUDENT AND RECENT GRADUATE COUNCIL INVITED SESSION: GETTING YOUR FIRST JOB

THE GRADUATE STUDENT AND RECENT GRADUATE COUNCIL

Victoria Liublinska*, Harvard University

The Graduate Student and Recent Graduate Council (GSRGC) to allow ENAR to better serve the special needs of students and recent graduates. I will describe the activities we envision the GSRGC participating in, how it would be constituted, and how it will interface with RAB and ENAR leadership.

email: vliublin@fas.harvard.edu

FINDING A POST-DOCTORAL FELLOWSHIP OR A TENURE-TRACK JOB

Eric Bair*, University of North Carolina Center for Neurosensory Disorders

This talk is primarily intended for doctoral students in statistics and will discuss jobs for statisticians in academia and strategies for finding jobs in academia. It will include a discussion of the benefits and drawbacks of working in academia as well as the types of academic jobs that exist. It will also discuss strategies for finding job openings, preparing job applications, interviewing, negotiating job offers, and other tactics for obtaining a job offer in academia.

email: ebair@email.unc.edu

GETTING YOUR FIRST JOB IN THE FEDERAL GOVERNMENT

Lillian Lin*, Centers for Disease Control and Prevention

The federal government is the largest single employer of statisticians in the United States yet most statistics graduate students do not know which agencies hire statisticians and are not familiar with the federal hiring process. The presenter has managed a statistics group since 2002. She will introduce nomenclature, review the distribution of federal statistician positions, describe typical job responsibilities, and advise on the application process.

email: lel5@cdc.gov

FINDING YOUR FIRST INDUSTRY JOB
Ryan May*, EMMES Corporation

This talk will focus on the characteristics that hiring managers are looking for at a CRO. This will include key items to include in your CV, along with interviewing tips and pitfalls. I will also discuss resources for students when looking for job openings. Finally, as a relatively recent graduate I will touch on my personal experiences in the job search.

email: rmay@emmes.com

60. STATISTICAL THERAPIES FOR HIGH-THROUGHPUT COMPLEX MISSING DATA AND DATA WITH MEASUREMENT BIAS

THE POTENTIAL AND PERILS OF PREPROCESSING: STATISTICAL PRINCIPLES FOR HIGH-THROUGHPUT SCIENCE
Alexander W. Blocker, Harvard University
Xiao-Li Meng*, Harvard University

Preprocessing forms an oft-neglected foundation for a wide range of statistical analyses. However, it is rife with subtleties and pitfalls. Decisions made in preprocessing constrain all later analyses and are typically irreversible. Hence, data analysis becomes a collaborative endeavor by all parties involved in data collection, preprocessing and curation, and downstream inference. This is particularly relevant as we contend with huge volumes of data from high-throughput biology. The technologies driving this data explosion are subject to complex new forms of measurement error. Simultaneously, we are accumulating increasingly massive biological databases. As a result, preprocessing has become a more crucial (and potentially more dangerous) component of statistical analysis than ever before. In this talk, we propose a theoretical framework for the analysis of preprocessing under the banner of multiphase inference. We provide some initial theoretical foundations for this area, building upon previous work in multiple imputation. We motivate this foundation with problems from biology, illustrating multiphase pitfalls and potential solutions in common and cutting-edge settings. This work suggests that principled statistical methods can, with renewed foundations, rise to meet the challenge of massive data with massive complexity.

email: ablocker@gmail.com

MISSING GENOTYPE INFERENCE AND ASSOCIATION ANALYSIS OF RARE VARIANTS IN ADMIXED POPULATIONS
Yun Li*, University of North Carolina, Chapel Hill
Mingyao Li; University of Pennsylvania
Yi Liu, University of North Carolina, Chapel Hill
Xianyun Mao, University of Pennsylvania
Wei Wang, University of North Carolina, Chapel Hill

Recent studies suggest rare variants (RVs) play an important role in the etiology of complex traits and exert larger genetic effects than common variants. However, there are two challenges for the analysis of RVs. First sequencing based studies which can experimentally discover RVs and call genotypes are still cost prohibitive for large numbers of individuals. Second traditional single marker tests are underpowered for the analysis of RVs. There are solutions proposed recently for both, but for genetically rather homogeneous populations, which are not optimal or even not valid for admixed populations, where the complex linkage disequilibrium (LD) patterns due to recent admixture makes both issues more challenging. We propose efficient methods for missing genotype inference and association testing of RVs in admixed populations.

email: yunli@med.unc.edu

A PENALIZED EM ALGORITHM FOR MULTIVARIATE GAUSSIAN PARAMETER ESTIMATION WITH NON-IGNORABLE MISSING DATA
Lin S. Chen*, University of Chicago
Ross L. Prentice, Fred Hutchinson Cancer Research Center
Pei Wang, Fred Hutchinson Cancer Research Center

For many modern high-throughput technologies, such as mass spectrometry based proteomics studies, missing values arise at high rates and the missingness probabilities often depend on the values to be measured. In this paper, we propose a penalized EM algorithm incorporating missing-data mechanism (PEMM) for estimating the mean and covariance of multivariate Gaussian data with non-ignorable missing data patterns that applies whether $p < n$ or $p >= n$. We estimate the parameters by maximizing a class of penalized likelihoods, in which the missing-data mechanisms are explicitly modeled. We illustrate the performance of PEMM with simulated and real data examples.

email: lchen@health.bsd.uchicago.edu

MIXTURE MODELING OF RARE VARIANT ASSOCIATION
Charles Kooperberg*, Fred Hutchinson Cancer Research Center
Benjamin Logsdon, Fred Hutchinson Cancer Research Center
James Y. Dai, Fred Hutchinson Cancer Research Center

We propose a new methodology for inference in genetic associations with rare variants. The approach uses a mixture model that divides rare variants in those that are and those that are not associated with a phenotype. While many burden tests have been proposed to identify genes with a burden of rare variants associated with phenotype, they generally do not acknowledge that even when some of the rare variants are associated with the phenotype, others are not. This is both biologically and statistically relevant for mapping rare variation because previous work indicates that even among non-synonymous mutation likely only 15-20% are functional. Our approach specifically models the presence of both functional and neutral variants with a discrete mixture distribution. We show through simulations that in many situations our approach has greater or comparable power to other popular burden tests, while also identifying the subset of rare variants associated with phenotype. Our algorithm leverages a fast variational Bayes approximate inference methodology to scale to exome-wide analyses. To demonstrate the efficacy of our approach we analyze platelet count within the National Heart, Lung, and Blood Institute's Exome Sequencing Project.

email: clk@fhcrc.org

61. ADVANCES IN INFERENCE FOR STRUCTURED AND HIGH-DIMENSIONAL DATA

THE MARCIENKO-PASTUR LAW FOR TIME SERIES
Haoyang Liu, University of California, Davis
Debashis Paul, University of California, Davis
Alexander Aue*, University of California, Davis

This talk discusses results about the behavior of the empirical spectral distribution of the sample covariance matrix of high-dimensional time series that extend the classical Marcienko-Pastur law. Specifically, we consider p-dimensional linear processes of the form $X_t = Z_t + \sum_{l=1}^\infty A_l Z_{t-l}$, where $\{Z_t : t \in \mathbb{Z}\}$ is a sequence of p-dimensional real or complex-valued random vectors with independent, zero mean, unit variance entries, and the $p\times p$ coefficient matrices $\{A_l\}_{l=1}^\infty$ are simultaneously diagonalizable and $\sum_{l} l \| A_l \| < \infty$. We analyze the limiting behavior of the empirical distribution of the eigenvalues of $\frac{1{n} \sum_{t=1}^n X_t X_t^*$ when $p/n \to c \in (0,\infty)$. This problem is motivated by applications in finance and signal detection.

email: aaue@ucdavis.edu

GENERALIZED EXPONENTIAL PREDICTORS FOR TIME SERIES

Prabir Burman*, University of California, Davis
Lu Wang, University of California, Davis
Alexander Aue, University of California, Davis
Robert Shumway, University of California, Davis

Generalized exponential predictors are proposed and investigated for univariate and multivariate time series. In the univariate case, linear combination of exponential predictors is used for forecasting. In the bivariate or multivariate case, linear combination is taken of the exponential predictors of the response as well as the covariates series. It can be shown that the proposed forecast models are dense in the stationary class of series. Computationally, these procedures are quite simple to implement as the standard programs for multiple regression can be employed to build the forecasts. Empirical examples as well as simulation studies are used to demonstrate the effectiveness of the proposed methods.

email: pburman@ucdavis.edu

ON ESTIMATION OF SPARSE EIGENVECTORS IN HIGH DIMENSIONS

Boaz Nadler*, Weizmann Institute of Science

In this talk we'll discuss estimation of the population eigenvectors from a high dimensional sample covariance matrix, under a low-rank spiked model whose eigenvectors are assumed to be sparse. We present several models of sparsity, corresponding minimax rates and a procedure that attains these rates. We will also discuss some differences between L_0 and L_q sparsity for q>0, as well as some limitations of recently suggested SDP procedures.

email: boaz.nadler@weizmann.ac.il

SPECTRA OF RANDOM GRAPHS AND THE LIMITS OF COMMUNITY IDENTIFICATION

Raj Rao Nadakuditi*, University of Michigan

We study networks that display community structure -- groups of nodes within which connections are unusually dense. Using methods from random matrix theory, we calculate the spectra of such networks in the limit of large size, and hence demonstrate the presence of a phase transition in matrix methods for community detection, such as the popular modularity maximization method. The transition separates a regime in which such methods successfully detect the community structure from one in which the structure is present but is not detected. Comparing these results with recent analyses of maximum-likelihood methods suggests that spectral modularity maximization is an optimal detection method in the sense that no other method will succeed in the regime where the modularity method fails.

email: rajnrao@umich.edu

62. FUNCTIONAL NEUROIMAGING DECOMPOSITIONS

LARGE SCALE DECOMPOSITIONS FOR FUNCTIONAL IMAGING STUDIES

Brian S. Caffo*, Johns Hopkins Bloomberg School of Public Health
Ani Eloyan, Johns Hopkins Bloomberg School of Public Health
Juemin Yang, Johns Hopkins Bloomberg School of Public Health
Seonjoo Lee, Johns Hopkins Bloomberg School of Public Health
Shanshan Li, Johns Hopkins Bloomberg School of Public Health
Shaojie Chen, Johns Hopkins Bloomberg School of Public Health
Lei Huang, Johns Hopkins Bloomberg School of Public Health
Huitong Qiu, Johns Hopkins Bloomberg School of Public Health
Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

In this talk we consider large-scale outer product models for functional imaging data. We investigate various forms of outer product models and discuss their strengths and weaknesses with a special focus on scalability. We suggest methods for using model based scores as outcome predictors. We apply the methods to novel large scale resting state functional imaging studies including the study of normal aging and developmental disorders.

email: bcaffo@jhsph.edu

A BAYESIAN APPROACH FOR MATRIX DECOMPOSITIONS FOR NEUROIMAGING DATA

Ani Eloyan*, Johns Hopkins Bloomberg School of Public Health
Sujit K. Ghosh, North Carolina State University

With the increasing amount of data available from functional imaging the development of well-rounded matrix decomposition methods is of interest. Several matrix decomposition methods such as Singular Value Decomposition, Independent Component Analysis, etc. are widely used by the neuriscience practitioners and are available as part of several imaging software. However, a large gap still exists in the development of Bayesian methods for analysis of neuroimaging data mainly due to the sheer size of the data and computational difficulties. We present a Bayesian dimension reduction method in line with blind source separation for resting state fMRI data and discuss the pros and cons of using Bayesian modeling for large-scale datasets.

email: aeloyan@jhsph.edu

MODELING COVARIATE EFFECTS IN INDEPENDENT COMPONENT ANALYSIS OF fMRI DATA

Ying Guo*, Emory University Rollins School of Public Health
Ran Shi, Emory University Rollins School of Public Health

Independent component analysis (ICA) has become an important tool for identifying and characterizing brain functional networks in functional magnetic resonance imaging (fMRI) studies. A key interest in such studies is to understand between-subject variability in the distributed patterns of the functional networks and its association with relevant subject characteristics. With current ICA methods, subjects' covariate effects are mainly investigated in post-ICA secondary analyses but not taken into account in the ICA model itself. We propose a new statistical model for group ICA that can directly incorporate subjects' covariate effects in decomposition of multi-subject fMRI data. Our model provides a formal statistical method to examine whether and how spatial distributed patterns of functional networks vary among subjects with different demographical, biological and clinical characteristics. We will discuss estimation methods for the new group ICA model. Simulation studies and application to an fMRI data example would also be presented.

email: yguo2@emory.edu

DYNAMICS OF INTRINSIC BRAIN NETWORKS

Vince Calhoun*, The Mind Research Network and the University of New Mexico
Eswar Damaraju, The Mind Research Network
Elena Allen, University of Bergen, Norway

There has been considerable interest in the intrinsic brain networks extracted from resting or task-based fMRI data. However most of the focus has been on estimating static networks from the data. In this talk I motivate the importance of evaluating how the intrinsic network properties (e.g. temporal interactions and spatial patterns) change over time during the course of an experiment. I present an approach for capturing dynamics and show results from several large data sets including healthy individuals and patients with schizophrenia. Findings suggest a specific set of regions including the default mode network which show increased variability in their dynamic frequencies, a so-called zone of instability, which may be important for adaptation and shows impairment in the patient data.

email: vcalhoun@unm.edu

63. STATISTICAL METHODS FOR TRIALS WITH HIGH PLACEBO RESPONSE

BEYOND CURRENT ENRICHMENT DESIGNS USING PLACEBO NON-RESPONDERS

Yeh-Fong Chen*, U.S. Food and Drug Administration

Several designs based on the enriched population have been proposed to deal with high placebo response commonly seen in psychiatric trials. Among these are a design with the placebo lead-in phase, a sequential parallel design (Fava et al., 2003), and a two-way enriched clinical trial design (Ivanova and Tamura, 2011). One common feature of these designs is the focus on placebo non-responders in an attempt to eliminate the influence of placebo responders on the effect size. However, the onset of treatments' effect may vary by disease pathology, so an optimal duration for the placebo lead-in phase is uncertain, and in turn the effectiveness of these enrichment designs is debatable. In this presentation, we will share our evaluations of these enrichment designs and compared them with our proposed new design strategy.

email: yehfong.chen@fda.hhs.gov

COMPARING STRATEGIES FOR PLACEBO CONTROLLED TRIALS WITH ENRICHMENT

Anastasia Ivanova*, University of North Carolina, Chapel Hill

We describe several two-stage design strategies for a placebo controlled trial where treatment comparison in the second stage is performed in an enriched population. Examples include placebo lead-in, randomized withdrawal and sequential parallel comparison design. Using the framework of the recently proposed two-way enriched design which includes all of these strategies as special cases we give recommendation on which two-stage strategy to use. Robustness of various designs is discussed.

email: aivanova@bios.unc.edu

REDUCING EFFECT OF PLACEBO RESPONSE WITH SEQUENTIAL PARALLEL COMPARISON DESIGN FOR CONTINUOUS OUTCOMES

Michael J. Pencina*, Boston University and Harvard Clinical Research Institute
Gheorghe Doros, Boston University
Denis Rybin, Boston University
Maurizio Fava, Massachusetts General Hospital

The Sequential Parallel Comparison Design (SPCD) is a novel approach intending to limit the effect of high placebo response in clinical trials. It can be applied to studies with binary as well as ordinal or continuous outcomes. Analytic methods proposed to date for continuous data included methods based on seemingly unrelated regression and ordinary least squares. Both ignore some data in estimating the analytic model and have to rely on imputation techniques to account for missing data. To overcome these issues we propose a repeated measures linear mixed model which uses all outcome data collected in the trial and accounts for data that is missing at random. An appropriate contrast formulated based on the final model is used to test the primary hypothesis of no difference in treatment effects between study arms pooled across the two phases of the SPCD trial. Simulations show that our approach preserves the type I error even for small sample sizes and offers adequate power under a wide variety of assumptions.

email: mpencina@bu.edu

64. COMPOSITE/PSEUDO LIKELIHOOD METHODS AND APPLICATIONS

DOUBLY ROBUST PSEUDO-LIKELIHOOD ESTIMATION FOR INCOMPLETE DATA

Geert Molenberghs*, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium
Geert Verbeke, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium
Michael G. Kenward, London School of Hygiene and Tropical Medicine, UK
Birhanu Teshome Ayele, I-BioStat, Hasselt Universiteit & Katholieke Universiteit Leuven, Belgium

In applied statistical practice, incomplete measurement sequences are the rule rather than the exception. Fortunately, in a large variety of settings, the stochastic mechanism governing the incompleteness can be ignored without hampering inferences about the measurement process. While ignorability only requires the relatively general missing at random assumption for likelihood and Bayesian inferences, this result cannot be invoked when non-likelihood methods are used. A direct consequence of this is that a popular non-likelihood-based method, such as generalized estimating equations, needs to be adapted towards a weighted version or doubly-robust version, when a missing at random process operates. So far, no such modification has been devised for pseudo-likelihood based strategies. We propose a suite of corrections to the standard form of pseudo-likelihood, to ensure its validity under missingness at random. Our corrections follow both single and double robustness ideas, and is relatively simple to apply. When missingness is in the form of dropout in longitudinal data or incomplete clusters, such a structure can be exploited towards further corrections. The proposed method is applied to data from a clinical trial in onychomycosis and a developmental toxicity study.

email: geert.molenberghs@uhasselt.be

COMPOSITE LIKELIHOOD INFERENCE FOR COMPLEX EXTREMES

Emeric Thibaud*, Anthony Davison and Raphaël Huser
École polytechnique fédérale de Lausanne

Complex extreme events, such as heat-waves and flooding, have major effects on human populations and environmental sustainability, and there is growing interest in modelling them realistically. Marginal modeling of extremes, through the use of the generalized extreme-value and generalized Pareto distributions, is well-developed, but spatial theory, which relies on max-stable processes, is needed for more complex settings and is currently an active domain of research. Max-stable models have been proposed for various types of data, but unfortunately classical likelihood inference cannot be used with them, because only their pairwise marginal distributions can be calculated in general. The use of composite likelihood makes inference feasible for such complex processes. This talk will describe the major issues in such modelling, and illustrate them with an application to extreme rainfall in Switzerland.

email: e.thibaud@gmail.com

STANDARD ERROR ESTIMATION IN THE EM ALGORITHM WHEN JOINT MODELING OF SURVIVAL AND LONGITUDINAL DATA

Cong Xu, University of California, Davis
Paul Baines, University of California, Davis
Jane-Ling Wang*, University of California, Davis

Joint modeling of survival and longitudinal data has been studied extensively in recent literature. The likelihood approach is one of the most popular estimation methods employed within the joint modeling framework. Typically the parameters are estimated using maximum likelihood, with computation performed by the EM algorithm. However, one drawback of this approach is that standard error (SE) estimates are not automatically produced when using the EM algorithm. Many different procedures have been proposed to obtain the asymptotic variance-covariance matrix for the parameters when the number of parameters

is typically small. In the joint modeling context, however, there is an infinite dimensional parameter, the cumulative baseline hazard function, which makes the problem much more complicated. In this talk, we show how to extend several existing parametric methods for EM SE estimation to our semiparametric setting, evaluate their precision and computational speed and compare them with the profile likelihood method and bootstrap method.

email: janelwang@ucdavis.edu

COMPOSITE LIKELIHOOD APPROACH FOR REGIME-SWITCHING MODEL

Jiahua Chen*, University of British Columbia, Vancouver, Canada
Peiming Wang, Auckland University of Technology, New Zealand

We present a composite likelihood approach as an alternative to the full likelihood approach for the two-state Markov regime-switching model for exchange rate time series. The proposed method is based on the joint density of pairs of consecutive observations, and its asymptotic properties are discussed. A simulation study indicates that the proposed method is more efficient and has better in-sample performance. An analysis of the USD/GBP exchange rate shows that the proposed method is more robust for inference about changes in the exchange-rate regime and better than the full likelihood method in terms of both in-sample and out-of-sample performance.

email: jhchen@stat.ubc.ca

65. RECENT ADVANCES IN ASSESSMENT OF AGREEMENT FOR CLINICAL AND LAB DATA

UNIFIED AND COMPARATIVE MODELS ON ASSESSING AGREEMENT FOR CONTINUOUS AND CATEGORICAL DATA

Lawrence Lin*, JBS Consulting Services Company

This presentation will be focused on the convergence of agreement statistics for continuous, binary, and ordinal data. MSD= E(X-Y)^2 is proportionally related to the total deviation index (TDI) for normally distributed data, proportionally related to one minus the crude weighted agreement probability for ordinal data, and is one minus the crude agreement probability for binary data. MSD is equivalently (in both estimation and statistical inference) scaled into the concordance correlation coefficient (CCC) for continuous data, weighted kappa for ordinal data, and kappa for binary data. Such convergence allows us to form a unified approach in assessing agreement for continuous, ordinal, and binary data from the paired samples model to more complex models when we have multiple raters and each rater has replicates per sample. Here, intra-rater precision, inter-rater agreement based on the average of replicates, and total-rater agreement based on individual readings will be discussed for both un-scaled and scaled agreement coefficients. We also consider a flexible and general setting where the agreement of certain cases can be compared relative to the agreement of a chosen case, such as individual bioequivalence, and comparing precision across raters.

email: equeilin@gmail.com

THE INTERPRETATION OF THE INTRACLASS CORRELATION COEFFICIENT IN THE AGREEMENT ASSAY

Josep L. Carrasco*, University of Barcelona

The analysis of concordance among repeated measures has received a huge amount of attention in the statistical literature leading to a range of different approaches. Among them the intraclass correlation coefficient (ICC) is one of the most applied and earlier introduced. However the ICC has been criticized because the complexity to interpret the ICC in terms of the analyzed variable scale and its dependence on the covariance among measures (between-subjects variance). Furthermore, those approaches only based on the within-subjects differences, as the total deviation index (TDI), have been called pure agreement indices. The TDI is an appealing approach to assess concordance that is non-covariance dependent and its values are in the same scale that the analyzed variable. The TDI is defined as the boundary such that differences in paired measurements of each subject are within the boundary with some determined probability. Regardless which approach is used the conclusions about the degree of concordance should be similar. Nevertheless, here two examples will be introduced where the ICC and the TDI give contradictory results that help to better understand what is really expressing the ICC.

email: jlcarrasco@ub.edu

MEASURING AGREEMENT IN METHOD COMPARISON STUDIES WITH HETEROSCEDASTIC MEASUREMENTS

Lakshika Nawarathna, University of Texas, Dallas
Pankaj K. Choudhary*, University of Texas, Dallas

It is common to use a two-step approach to evaluate agreement between methods of measurement in a method comparison study. In the first step, one fits a suitable mixed model to the data. This fitted model is then used in the second step to perform inference on agreement measures, e.g., concordance correlation and total deviation index, which are functions of the parameters in the model. However, a frequent violation of the standard mixed model assumptions is that the error variability may depend on the unobservable magnitude of measurement. If this heteroscedasticity is not taken into account, the resulting agreement evaluation may be misleading as the extent of agreement between the methods is not constant anymore. To deal with this situation, we consider a modification of the common two-step approach wherein a heteroscedastic mixed model is used in the first step. The second step remains the same as before except that now confidence bands for agreement measures are used for agreement evaluation. This methodology is illustrated by applying it to a cholesterol data set from the literature. The results of a simulation study are also presented.

email: pankaj@utdallas.edu

AN AUC-LIKE INDEX FOR AGREEMENT ASSESSMENT

Zheng Zhang*, Brown University
Youdan Wang, Brown University
Fenghai Duan, Brown University

The commonly used statistical measures for assessing agreement of readings generated by multiple observers or raters, such as intraclass correlation coefficient (ICC) or concordance correlation coefficient (CCC), have well-known dependency on the data's normality assumption, hereby are heavily influenced by data outliers. Here we propose a novel agreement measure (rank-based agreement index, rAI) by estimating agreement from data's overall ranks. Such non-parametric approach provides a global measure of agreement, regardless of data's exact distributional form. We have shown rAI as a function of the overall ranking of each subject's extreme values. Furthermore, we propose an agreement curve, a graphic tool that aids visualizing extent of the agreement, which strongly resembles the receiver operating characteristic (ROC) curve. We further show rAI is a function of the area under the agreement curve. Consequently, rAI shares some important features with the area under of the ROC curve (AUC). Extensive simulation studies are included. We illustrate our method with two cancer imaging study datasets.

email: zzhang@stat.brown.edu

66. FUNCTIONAL DATA ANALYSIS

TESTING THE EFFECT OF FUNCTIONAL COVARIATE FOR FUNCTIONAL LINEAR MODEL

Dehan Kong*, North Carolina State University
Ana-Maria Staicu, North Carolina State University
Arnab Maity, North Carolina State University

In this article, we consider the functional linear model with a scalar response. Our goal is to test for no effect of the model, that is to test whether the functional coefficient function equals zero. We use the functional principal component analysis and write the functional linear model as a linear combination of the functional principal component scores. Various traditional tests such as Wald, score, likelihood ratio and F test are applied. We compare the performance of these tests under both regular dense design and sparse irregular design. We also do some research on how sample size affect the performance of these tests. Both the asymptotic null distribution and the alternative distribution are derived for those tests that work well. We have also discussed about sample size needed to achieve certain power. We demonstrate our results using simulations and real data.

email: dkong2@ncsu.edu

ACCELEROMETRY METRICS FOR EPIDEMIOLOGY

Jiawei Bai*, Johns Hopkins University
Bing He, Johns Hopkins University
Thomas A. Glass, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University

We introduce a set of metrics for human activity based on high density acceleration recordings from a hip worn three-axis accelerometer. Data were collected from 34 older subjects who wore the devices for up to seven days during their daily living activities. We propose simple metrics that are based on two concepts: 1) time active, a measure of the length of time when the subject activity is distinguishable from rest; and 2) activity intensity, a measure of relative amplitude of activity relative to rest. Both measurements are time dependent, but their means and standard deviations are reasonable and complementary summaries of daily activity. All measurements are normalized (have the same interpretation across subjects and days), easy to explain and implement, and reproducible across platforms and software implementations. This is a non-trivial task in an observational study where raw acceleration can be dramatically affected by the location of the device, angle with respect to body, body geometry, subject-specific size and direction of energy produced, time, battery voltage, and other unpredictable, but often occurring, events. The results of a small study of the association between our activity metrics and health outcomes shows promising initial results.

email: jbai@jhsph.edu

SPARSE SEMIPARAMETRIC NONLINEAR MODEL WITH APPLICATION TO CHROMATOGRAPHIC FINGERPRINTS

Michael R. Wierzbicki*, University of Pennsylvania
Li-bing Guo, Guangdong College of Pharmacy
Qing-tao Du, Guangdong College of Pharmacy
Wensheng Guo, University of Pennsylvania

Chromatography is a popular tool in determining the chemical composition of biological samples. For example, traditional Chinese herbal medications are comprised of numerous compounds and identifying commonalities across a set of samples is of interest in quality control and identification of active compounds. Chromatographic experiments output a plot of all the detected abundances of the compounds over time. The resulting chromatogram is characterized by a number of sharp spikes each of which corresponds to the presence of a different compound. Due to variation in experimental conditions, a given spike is often not aligned in time across a set of samples. We propose a sparse semiparametric nonlinear model for the establishment of a standardized chromatographic fingerprint from a set of chromatograms under different experimental conditions. Our framework results in simultaneous alignment, model selection, and estimation of chromatograms. Wavelet basis expansion is used to model the common shape function of the curves nonparametrically. Curve registration is performed by parametric modeling of the time transformations. Penalized likelihood with the adaptive lasso penalty provides a unified criterion for model selection and estimation. The adaptive lasso estimators are shown to possess the oracle property. We apply the model to data of the medicinal plant, rhubarb.

email: mwierz@mail.med.upenn.edu

REGULARIZED 3D FUNCTIONAL REGRESSION FOR BRAIN IMAGING VIA HAAR WAVELETS

Xuejing Wang*, University of Michigan
Bin Nan, University of Michigan
Ji Zhu, University of Michigan
Robert Koeppe, University of Michigan

There has been an increasing interest in the analysis of functional data in recent years. Samples of curves, images, or other functional observations are often collected in many fields. Our primary motivation and application come from brain imaging studies on cognitive impairment in elderly subjects with brain disorders. We propose a highly effective regularized Haar-wavelet-based approach for the analysis of three-dimensional brain imaging data in the framework of functional data analysis, which automatically takes into account the spatial information among neighboring voxels. We conduct extensive simulation studies to evaluate the prediction performance of the proposed approach and its ability to identify related regions to the response variable, with the underlying assumption that only a few relatively small subregions are associated with the response variable. We then apply the proposed approach to search for brain subregions that are associated with cognitive impairment using PET imaging data.

email: xuejwang@umich.edu

MECHANISTIC HIERARCHICAL GAUSSIAN PROCESSES

Matthew W. Wheeler*, The National Institute for Occupational Safety and Health and University of North Carolina, Chapel Hill
David B. Dunson, Duke University
Amy H. Herring, University of North Carolina, Chapel Hill
Sudha P. Pandalai, The National Institute for Occupational Safety and Health
Brent A. Baker, The National Institute for Occupational Safety and Health

The statistics literature on functional data analysis focuses primarily on flexible black-box approaches, which are designed to allow individual curves to have essentially any shape while characterizing variability. Such methods typically cannot incorporate mechanistic information, which is commonly expressed in terms of differential equations. Motivated by studies of muscle activation, we propose a nonparametric Bayesian approach that takes into account mechanistic understanding of muscle physiology. A novel class of hierarchical Gaussian processes is defined that favors curves consistent with differential equations defined on motor, damper, spring systems. A Gibbs sampler is proposed to sample from the posterior distribution and applied to a study of rats exposed to non-injurious muscle activation protocols. Although motivated by muscle force data, a parallel approach can be used to include mechanistic information in broad functional data analysis applications.

email: mwheeler@cdc.gov

VARIABILITY ANALYSIS ON REPEATABILITY EXPERIMENT OF FLUORESCENCE SPECTROSCOPY DEVICES

Lu Wang*, Rice University
Dennis D. Cox, Rice University

This project is about to investigate the use of spectroscopic devices to detect cancerous and pre-cancerous lesions. One major problem with bio-medical applications of optical spectroscopy is the repeatability of the measurements. The measured spectra cannot be accurately measured; they are functional data with variations in

peak height between different devices and probes. This project identifies the variations and eliminates them from the measurement data. Iterative local quadratic model and functional principle component analysis method are developed here.

email: lw7@rice.edu

67. PERSONALIZED MEDICINE

HYPOTHESIS TESTING FOR PERSONALIZING TREATMENT

Huitian Lei*, University of Michigan
Susan Murphy, University of Michigan

In personalized/stratified treatment the recommended treatment is based on patient characteristics. We define a biomarker as useful in personalized decision making if for a particular value of the biomarker, there is sufficient evidence to recommend one treatment, while for other values of the biomarker, either there is sufficient evidence to recommend a different treatment, or there is insufficient evidence to recommend a particular treatment. We propose a two-stage hypothesis testing procedure for use in evaluating if a biomarker is useful in personalized decision making. In the first stage of the procedure, the sample space is partitioned based on the observed value of a test statistic for testing treatment-biomarker interaction. In the second stage of the procedure, the treatment effect for each value of the biomarker is tested; in this stage we control a conditional error rate. We illustrate the proposed procedure based using data from a depression study involving the medication, Nefazodone and a combination of a behavorial therapy with Nefazodone.

email: ehlei@umich.edu

NON-PARAMETRIC INFERENCE OF CUMULATIVE INCIDENCE FUNCTION FOR DYNAMIC TREATMENT REGIMES UNDER TWO-STAGE RANDOMIZATION

Idil Yavuz*, University of Pittsburgh
Yu Cheng, University of Pittsburgh
Abdus S. Wahed, University of Pittsburgh

Recently personalized medicine and dynamic treatment regimes have drawn considerable attention. Also, more and more practitioners become aware of competing-risk censoring for event type outcomes. We aim to compare several treatment regimes from a two-stage randomized trial on survival outcomes that are subject to competing-risk censoring. With the presence of competing risks, cumulative incidence function (CIF) has been widely used to quantify the probability of occurrence of the event of interest at or before a specific time point. However, if we only use the data from those subjects who have followed a specific treatment regime to estimate the CIF, the resulting estimator may be biased. Hence, we propose alternative non-parametric estimators for the CIF using inverse weighting, and provide inference procedures for the proposed estimator based on the asymptotic linear representation in addition to test procedures to compare the CIFs from two different treatment regimes. Through simulation we show the advantages of the proposed estimators compared to the standard estimator. Since dynamic treatment regimes are widely used in treating diseases that require complex treatment and competing-risk censoring is common in studies with multiple end-points, the proposed methods provide useful inferential tools that will help advocate research in personalized medicine.

email: idy1@pitt.edu

USING PSEUDO-OBSERVATIONS TO ESTIMATE DYNAMIC MARGINAL STRUCTURAL MODELS WITH RIGHT CENSORED RESPONSES

David M. Vock*, University of Minnesota
Anastasios A. Tsiatis, North Carolina State University
Marie Davidian, North Carolina State University

A dynamic treatment regime takes an individual's characteristics and clinical and treatment histories and dictates a particular treatment or sequence of treatments to receive. Dynamic marginal structural models have been proposed to estimate the average response had, contrary to fact, all subjects in the population followed a particular dynamic treatment regime within a given class of regimes. An added complication occurs when the response may be right censored. In this case, inverse probability of censoring weighted (IPCW) estimators may be used to obtain consistent estimators of the parameters in the marginal structural model. However, the data analyst must know or correctly specify a model for the censoring mechanism. Rather than use IPCW estimators, we show that under certain realistic assumptions how jackknife pseudo-observations may be used in place of the, potentially unobserved, failure time to derive consistent estimators of the parameters in the marginal structural model. We use our method to estimate the restricted 5-year mean survival of patients awaiting lung transplantation if, contrary to fact, all patients were to delay transplantation until their lung allocation score, a composite score of waitlist prognosis, surpassed certain thresholds.

email: vock@umn.edu

DOUBLE ROBUST ESTIMATION OF INDIVIDUALIZED TREATMENT FOR CENSORED OUTCOME

Yingqi Zhao*, University of Wisconsin, Madison
Donglin Zeng, University of North Carolina, Chapel Hill
Michael R. Kosorok, University of North Carolina, Chapel Hill

It is common to have heterogeneity among patients' responses to different treatments. Specifically, when the clinical outcome of interest is survival time, the goal is to maximize the expected time of survival by providing the right patient with the right treatments. We develop methodology for finding the optimal individualized treatment rules in the framework of censored data. Instead of regression modeling, we directly maximize the value function, i.e., the expected outcome for the specific rules, using non parametric methods. Moreover, to adjust for the censored data, we propose a double robust solution by estimating both survival and censoring time with two working models. When either model is correct, we obtain unbiased optimal decision rules. We conduct simulations which show the superior performances of the proposed method.

email: yqzhao@biostat.wisc.edu

PENALIZED REGRESSION AND RISK PREDICTION IN GENOME-WIDE ASSOCIATION STUDIES

Erin Austin*, University of Minnesota
Wei Pan, University of Minnesota
Xiaotong Shen, University of Minnesota

An important task in personalized medicine is to predict disease risk based on a person's genome, e.g. on a large number of single-nucleotide polymorphisms (SNPs). Genome-wide association studies (GWAS) make SNP and phenotype data available to researchers. A critical question for researchers is how to best predict disease risk. Penalized regression equipped with variable selection, such as LASSO, SCAD, and TLP, is deemed to be promising in this setting. However, the sparsity assumption taken by the LASSO, SCAD, TLP and many other penalized regression techniques may not be applicable here: it is now hypothesized that many common diseases are associated with many SNPs with small to moderate effects. In this project, we use the GWAS data from the Wellcome Trust Case Control Consortium (WTCCC) to investigate the performance of various unpenalized and penalized regression approaches under true sparse or non-sparse models. We find that in general penalized regression outperformed unpenalized regression; SCAD, LASSO, TLP, and elastic net weighted towards LASSO performed best for sparse models, while ridge regression and elastic net were the winners for non-sparse models; across all the scenarios, LASSO always worked well with its performance either equal or close to that of the winners.

email: austi260@umn.edu

TIME-SENSITIVE PREDICTION RULES FOR DISEASE RISK OR ONSET THROUGH LOCALIZED KERNEL MACHINE LEARNING

Tianle Chen*, Columbia University
Huaihou Chen, New York University
Yuanjia Wang, Columbia University
Donglin Zeng, University of North Carolina at Chapel Hill

Accurately classifying whether a presymptomatic subject is at risk of a disease using genetic and clinical markers offers the potential for early intervention well before the onset of clinical symptoms. For many diseases, the risk varies with age and varies with markers that are themselves dependent on age. This work aims at identifying effective time-sensitive classification and prediction rules for age-dependent disease risk or onset using age-sensitive biomarkers through localized kernel machine learning. In particular, we develop a large-margin classifier implemented with localized kernel support vector machine. We study the convergence rate of the developed rules as a function of kernel bandwidth and offer guidelines on the choice of the bandwidth as a function of the sample size. Ranking the biomarkers based on their cumulative effects provides an opportunity to select important markers. We extend our approach to longitudinal data through the use of a nonparametric decision function with random effects, where we model the main effects of biomarkers, time, and their interaction through appropriate kernel functions. Subject-specific random intercepts and random slopes are included in the decision function to account for patient heterogeneity and improve prediction. We apply the proposed methods to a real world Huntington's disease data.

email: tc2411@columbia.edu

IMPROVEMENTS TO THE INTERACTION TREES ALGORITHM FOR SUBGROUP ANALYSIS IN CLINICAL TRIALS

Yi-Fan Chen*, University of Pittsburgh
Lisa A. Weissfeld, University of Pittsburgh

With the advent of personalized medicine, the goal of an analysis is to identify subgroups that will receive the greatest benefit from a given treatment. The approach considered here centers on methods for clinical trials that are geared towards exploring heterogeneity between subjects and its impact on treatment response. One such approach is an extension of classification and regression trees (CART) based on the development of interaction trees so that subgroups within treatment arms are better identified. With these analyses it is possible to generate hypotheses for future clinical trials and to present the results in a readily accessible format. One major issue with this approach is the greediness of the algorithm and the difficulty of addressing it without losing the interpretability. We focus on this issue by integrating random forests and the evolutionary algorithm into the interaction trees algorithm, while preserving the tree structure. The advantage of this approach is that it allows for the identification of subgroups that benefit most from the treatment. We evaluate the properties of the modified interaction trees algorithm and compare it with the original interaction trees algorithm via simulations. The strengths of the proposed method are demonstrated through a survival data example.

email: yic33@pitt.edu

68. EPIDEMIOLOGIC METHODS IN SURVIVAL ANALYSIS

MATCHING IN THE PRESENCE OF MISSING DATA IN TIME-TO-EVENT STUDIES

Ruta Brazauskas*, Medical College of Wisconsin
Mei-Jie Zhang, Medical College of Wisconsin
Brent R. Logan, Medical College of Wisconsin

Matched pair studies in survival analysis are often done to examine the survival experience of patients with rare conditions or in the situation where extensive collection of additional information on cases and/or controls is required. Once matching is done, analysis is usually performed by using stratified or marginal Cox proportional hazards models. However, in many studies some patients will have missing values of the covariates they should be matched on. In this presentation, we will examine several methods that could be used to match cases and controls when some covariate values are missing. They range from matching using only individuals with complete data to more complex matching procedures performed after the imputation of the missing values of the covariates. A simulation study is used to explore the performance of these matching techniques under several patterns and proportions of missing data.

email: ruta@mcw.edu

APPLICATION OF TIME-DEPENDENT COVARIATES COX MODEL IN EXAMINING THE DYNAMIC ASSOCIATIONS OF BODY MASS INDEX AND CAUSE-SPECIFIC MORTALITIES

Jianghua He, University of Kansas Medical Center
Huiquan Zhang*, University of Kansas Medical Center

Previous studies have shown that the association of body mass index (BMI) and all-cause mortality is dynamic that different study designs may lead to different or even opposite results. To better understand the controversial association of BMI and all-cause mortality, this study is conducted to examine the association of BMI and cause-specific mortalities based on a pooled data with 33,144 individuals. BMI was transformed to capture the curvature association of BMI and mortality. The associations were analyzed using Cox model at first and the proportional hazards (PH) assumptions were tested. Time-dependent covariates Cox model was used to model the dynamic associations when the PH assumptions for any BMI related term were violated. Three specific types of causes of mortalities were examined, including cardiovascular disease (CVD), cancer, and other causes. For women, no dynamic association of BMI and cause-specific mortality was found. For men, the associations of BMI and three cause-specific mortalities were all dynamic. Time-dependent covariate Cox models were used to show that the dynamic associations of BMI with CVD, cancer and other causes mortalities were different.

email: winnie.huiquanzhang@gmail.com

GENERALIZED CASE-COHORT STUDIES WITH MULTIPLE EVENTS

Soyoung Kim*, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill

Case-cohort studies have been recommended for infrequent diseases or events in large epidemiologic studies. This study design consists of a random sample of the entire cohort, named the subcohort, and all the subjects with the disease of interest. When the rate of disease is not low or the number of cases are not small, the generalized case-cohort study which selects subset of all cases is used. When several diseases are of interest, several generalized casecohort studies are usually conducted using the same subcohort. The common practice is to analyze each disease separately ignoring data collected on sampled subjects with the other diseases. This is not an efficient use of the data. In this paper, we propose efficient estimation for proportional hazards model by making full use of available covariate information for the other diseases. We consider both joint analysis and separate analysis for the multiple diseases. We propose an estimating equation approach with a new weight function. We establish that the proposed estimator is consistent and asymptotically normally distributed. Simulation studies show that the proposed methods using all available information gain efficiency. We apply our proposed method to the data from the Busselton Health Study.

email: kimso@live.unc.edu

AN ORNSTEIN-UHLENBECK RANDOM EFFECTS THRESHOLD REGRESSION CURE RATE MODEL

Roger A. Erich; Air Force Institute of Technology
Michael L. Pennell*, The Ohio State University

In cancer clinical trials, researchers typically examine the effects of a treatment on progression-free or relapse-free survival. If effective, the treatment results in a fraction (p) of subjects who will not experience a tumor recurrence (i.e., are cured). In this presentation, we present a cure rate model for time to event data which models patient health using a latent stochastic process with an event occurring once the process reaches a predetermined threshold for the first time. In our model, a patient's health follows one of two different processes: with probability p, the patient's health remains stable while the health of the remaining (not cured) patients follows an Ornstein-Uhlenbeck process which gravitates toward the threshold which triggers the event. The initial state of each subject's process is related to observed and unobserved covariates, the latter being accommodated through the use of a gamma distributed random effect. A logistic regression model is used to relate the cure rate (p) to observed covariates. In addition to being a conceptually appealing model, our approach avoids the proportional hazards assumption of some commonly used cure rate models. We demonstrate our approach using relapse-free survival data of high-risk melanoma patients undergoing definitive surgery.

email: mpennell@cph.osu.edu

ACCOUNTING FOR LENGTH-BIAS AND SELECTION BIAS IN ESTIMATING MENSTRUAL CYCLE LENGTH

Kirsten J. Lum*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health and Johns Hopkins Bloomberg Schoolof Public Health
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health

Prospective pregnancy studies are a valuable source of longitudinal data on menstrual cycle length. However, care is needed when making inferences. For example, accounting for the sampling plan is necessary for unbiased estimation of the menstrual cycle length distribution. If couples can enroll when they learn of the study as opposed to waiting for the next menstrual cycle, then due to length-bias, the enrollment cycle will be stochastically larger than the general run of cycles, a typical property of prevalent cohort studies. Furthermore, the probability of enrollment can depend on the length of time since a woman s last menstrual period (a backward recurrence time), resulting in a form of selection bias. We propose an estimation procedure which accounts for length-bias and selection bias in the likelihood for enrollment menstrual cycle length. Simulation studies quantify performance when proper account is taken of these features and when the likelihood fails to account for them. In addition, we illustrate the approach using data from the Longitudinal Investigation of Fertility and the Environment Study (LIFE Study).

email: kirsten.lum@gmail.com

INCORPORATING A REPRODUCIBILITY SAMPLE INTO MULTI-STATE MODELS FOR INCIDENCE, PROGRESSION AND REGRESSION OF AGE-RELATED MACULAR DEGENERATION

Ronald E. Gangnon*, University of Wisconsin
Kristine E. Lee, University of Wisconsin
Barbara EK Klein, University of Wisconsin
Sudha K. Iyengar, Case Western Reserve University
Theru A. Sivakumaran, Cincinnati Children's Medical Center
Ronald Klein, University of Wisconsin

Misclassification of disease status is a common problem in multi-state models of disease status for panel data. For example, gradings of age-related macular degeneration (AMD) severity are subject to misclassification from a variety of sources (subjects, photographs, graders). In the Beaver Dam Eye Study (BDES), AMD status on a 5-level severity scale was graded from retinal photographs taken at up to 5 study visits between 1988 and 2010 for 4,379 persons aged 43 to 86 years at the time of initial examination. The BDES includes a reproducibility sample of 89 photographs with 17 to 84 (median 20) gradings per photograph. We consider 5 methods for incorporating information from the reproducibility sample into a multi-state model for incidence, progression and regression of AMD: (1) without using the reproducibility sample, (2) using point estimates of the misclassification matrix from the reproducibility sample, (3) using a parametric bootstrap sample of the misclassification matrix from the reproducibility sample, (4) using a non-parametric bootstrap sample of the misclassification matrix from the reproducibility sample, and (5) using the joint likelihood for the longitudinal and reproducibility samples. We compare the methods in terms of inferential results and computational burden.

email: ronald@biostat.wisc.edu

69. POWER AND SAMPLE SIZE

DECISION RULES AND ASSOCIATED SAMPLE SIZE PLANNING FOR REGIONAL APPROVAL IN MULTIREGIONAL CLINICAL TRIALS

Rajesh Nair*, U.S. Food and Drug Administration
Nelson Lu, U.S. Food and Drug Administration
Yunling Xu, U.S. Food and Drug Administration

Recently, multi-regional trials have been gaining momentum around the globe and many medical product manufacturers are now eager to conduct multi-region/country trials with the purpose of gaining regulatory approval from multiple-regions/countries simultaneously. Such a strategy has the potential to make safe and effective medical products more quickly available to patients globally. As regulatory decisions are always made in a local context, this also poses huge regulatory challenges. We will discuss two conditional decision rules which can be used for medical product approval by local regulatory agencies based on the results of a multi-regional clinical trial. We also illustrate sample size planning for one-arm and two-arm trials with binary endpoint.

email: rajesh.nair@fda.hhs.gov

MULTI-REGIONAL CLINICAL TRIAL DESIGN AND CONSISTENCY ASSESSMENT OF TREATMENT EFFECTS

Hui Quan, Sanofi
Xuezhou Mao*, Sanofi and Columbia University, Mailman School of Public Health
Joshua Chen, Merck Research Laboratories
Weichung Joe Shih, University of Medicine and Dentistry of New Jersey
Soo Peter Ouyang, Celgene Corporation
Ji Zhang, Sanofi
Peng-Liang Zhao, Sanofi
Bruce Binkowitz, Merck Research Laboratories

For multi-regional clinical trial design and data analysis, fixed effect and random effect models can be applied. Thoroughly understanding the features of these models in a MRCT setting will help us to assess the applicability of a model for a MRCT. In this paper, the interpretations of trial results from these models are discussed. The impact of the number of regions and the sample size configuration across the regions on the required total sample size, the estimation of between-region variability and type I error rate control for the overall treatment effect assessment is also evaluated. For estimating the treatment effects of individual regions, the empirical shrinkage estimator and the James-Stein type shrinkage estimator associate with a smaller variability compared to the regular sample mean estimator. Computation and simulation are conducted to compare the performance of these estimators when they are applied to assess consistency of treatment effects across regions. A multinational trial example is used to illustrate the application of the methods.

email: xm2126@columbia.edu

SAMPLE SIZE/POWER CALCULATION FOR STRATIFIED CASE-COHORT DESIGN

Wenrong Hu*, University of Memphis
Jianwen Cai, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill

Case-cohort study design (CC) has been often used for the risk factor assessment in epidemiologic studies or disease prevention trials in situations when the disease is rare. Sample size/power calculation for CC is given in

Cai and Zeng (2004). However, the sample size/power calculation for stratified case-cohort (SCC) design has not been addressed before. This article extends the results in Cai and Zeng (2004) to SCC by introducing the stratified log-rank type test statistic for the SCC design and deriving the sample size/power calculation formula. Simulation studies show that the proposed test for SCC with small sub-cohort sampling fractions is valid and efficient for the situations where the disease rate is low. Furthermore, optimization of sampling in SCC is discussed in comparison with the proportional and balanced sampling techniques, and the corresponding sample size calculation formulas are derived for these three designs. Theoretical powers from these three designs are compared for the situations when the disease rates are homogeneous and heterogeneous over the strata. The results show that either the proportional or balanced design can possibly yield higher power than the other. The optimal design yields the highest power with the smallest required sample size among all three designs.

email: wenrong.hu@cslbehring.com

THE EFFECT OF INTERIM SAMPLE SIZE RECALCULATION ON TYPE I AND II ERRORS WHEN TESTING A HYPOTHESIS ON REGRESSION COEFFICIENTS

Sergey Tarima*, Medical College of Wisconsin
Aniko Szabo, Medical College of Wisconsin
Peng He, Medical College of Wisconsin
Tao Wang, Medical College of Wisconsin

The dependence of sample size formulas on nuisance parameters inevitably forces investigators to use estimates of these nuisance parameters. We investigated the effect of naive interim sample size recalculation based on re-estimated nuisance parameters on type I and II errors. More than 200 simulation studies with 10,000 Monte-Carlo repetitions each were completed covering various scenarios: linear and logistic regressions; two to ten predictors; binary and continuous predictors; no, moderate, and strong correlation between predictors; different treatment effects; different internal pilot sample sizes and different upper bounds for the total sample size. Only Wald tests were considered in our investigation. For linear models we observed a minor positive inflation of the type I error increasing as the expected sample size getting closer to the sample size of the internal pilot. We also noted that power increases for these cases. If the expected sample size is getting closer to the upper bound of the total sample size, power decreases. Internal sample size recalculation for logistic regression models often produces too conservative type I errors and/or overpowered study designs.

email: sergey.s.tarima@gmail.com

A COMMENT ON SAMPLE SIZE CALCULATIONS FOR BINOMIAL CONFIDENCE INTERVALS

Lai Wei*, State University of New York at Buffalo
Alan D. Hutson, State University of New York at Buffalo

We firstly examine sample size calculations for a binomial proportion based on the confidence interval width of the Agresti-Coull, Wald and Wilson Score intervals. We point out that the commonly used methods based on known and fixed standard errors cannot guarantee the desired confidence interval width given a hypothesized proportion. Therefore, a new adjusted sample size calculation method is introduced, which is based on the conditional expectation of the width of the confidence interval given the hypothesized proportion. This idea is further extended to Poisson distribution and continuous distribution, such as exponential distribution.

email: laiwei@buffalo.edu

OPTIMAL DESIGN FOR DIAGNOSTIC ACCURACY STUDIES WHEN THE BIOMARKER IS SUBJECT TO MEASUREMENT ERROR

Matthew T. White*, Boston Children's Hospital
Sharon X. Xie, University of Pennsylvania

Biomarkers have become increasingly important in recent years for their ability to effectively distinguish diseased from non-diseased individuals, potentially avoiding unnecessary and invasive diagnostic testing. Given their importance, it is critical to design and analyze studies to produce reliable estimates of diagnostic performance. Biomarkers are often obtained with measurement error, which may cause the biomarker to appear ineffective if not taken into account in the analysis. We develop optimal design strategies for studying the effectiveness of an error prone biomarker in differentiating diseased from non-diseased individuals and focus on the area under the receiver operating characteristic curve (AUC) as the primary measure of effectiveness. Using an internal reliability sample within the diseased and non-diseased groups, we develop optimal study design strategies that 1) minimize the variance of the estimated AUC subject to constraints on the total number of observations or total cost of the study or 2) achieve a pre-specified power. We develop optimal allocations of the number of subjects in each group, the size of the reliability sample in each group, and the number of replicate observations per subject in the reliability sample in each group under a variety of commonly seen study conditions.

email: matthew.thomas.white@gmail.com

STUDY DESIGN IN THE PRESENCE OF ERROR-PRONE SELF-REPORTED OUTCOMES

Xiangdong Gu*, University of Massachusetts, Amherst
Raji Balasubramanian, University of Massachusetts, Amherst

In this paper, we consider data settings in which the outcome of interest is a time-to-event random variable that is observed at intermittent time points through error-prone tests. For this setting, using a likelihood-based approach we develop methods for power and sample size calculations and compare the relative efficiency between perfect and imperfect diagnostic tests. Our work is motivated by diabetes self-reports among the approximately 160,000 women enrolled in the Women's Health Initiative. We also compare designs under different missing data mechanisms and evaluate the effect of error-prone disease ascertainment at baseline.

email: xdgu@schoolph.umass.edu

70. MULTIPLE TESTING

MULTIPLICITY ADJUSTMENT OF MULTI-LEVEL HYPOTHESIS TESTING IN IMAGING BIOMARKER RESEARCH

Shubing Wang*, Merck

In imaging biomarker research, longitudinal studies are often used for reproducibility and efficacy validation. The multiple biomarkers, multiple groups and repeated measures impose the multiple testing of biomarker efficacies complex multiple-level structures. Less powerful conventional multiplicity adjustment methods, either have no explicit assumption of correlation structures, such as False Discovery Rate (FDR), or only assume simple one-dimensional correlations, such as Random Field Theory. The author proposed a multi-level multiplicity adjustment method by first building the hierarchical structures of multiple tests. On the highest level are the biomarkers, which are usually correlated, but not ordered. Within a biomarker, the highly correlated tests, such as within-group change tests, are first identified. The joint distribution of these tests can be calculated based on a linear mixed-effects model. Therefore, the explicit distribution of the simultaneous confidence band can be estimated, as well as the adjusted p-values. The rest of the tests belong to the less correlated test category. We then apply a double FDR procedure proposed by Mehrotra and Heyse to set up the different thresholds and to calculate adjusted p-values at different levels.

email: shubing_wang@merck.com

MULTIPLICITY STRATEGIES FOR MULTIPLE TREATMENTS AND MULTIPLE ENDPOINTS

Kenneth Liu*, Merck
Paulette Ceesay, Merck
Ivan Chan, Merck
Nancy Liu, Merck
Duane Snavely, Merck
Jin Xu, Merck

When testing multiple hypotheses, multiplicity adjustments are used to control the Type 1 error. Many clinical trials test multiple hypotheses which are organized into multiple treatments and multiple endpoints. We present a framework that provides multiplicity strategies for these types of problems. In particular, graphical solutions (Bretz et al, 2009) and shortcuts (Hommel et al, 2007) will be presented using examples.

email: Kenneth_Liu@Merck.com

MULTIPLE COMPARISONS WITH THE BEST FOR SURVIVAL DATA WITH TREATMENT SELECTION ADJUSTMENT

Hong Zhu*, The Ohio State University
Bo Lu, The Ohio State University

Many clinical trials and observational studies have time to some event as their endpoint, and it is often interesting to compare several treatments in terms of their survival curves. In some situations, not all pairwise comparisons are necessary and the primary focus is comparisons with the unknown best treatment. Methods of multiple comparison with the best (MCB) have been developed for and applied in linear and generalized linear model setting to provide simultaneous confidence intervals for the difference between each treatment and the best of the others (Hsu, 1996). However, comparing several groups in terms of their survival outcomes has not yet received much attention. We focus on MCB for survival data under random right censoring, which identifies the treatments with practically maximum benefit. In addition, in an observation study or a non-randomized clinical trial, the sample in different groups may be biased due to confounding effects. Cox model incorporating potential confounders as covariates typically allows for adjustment, but in some cases, the proportionality assumption may be invalid. We extend the method of MCB to survival data, and propose a new procedure based on a stratified Cox model, using propensity score stratification to reduce confounding bias. A motivating example is discussed for illustration of the method.

email: hzhu@cph.osu.edu

AN ADAPTIVE RESAMPLING TEST FOR DETECTING THE PRESENCE OF SIGNIFICANT PREDICTORS

Ian McKeague, Columbia University
Min Qian*, Columbia University

This paper constructs a screening procedure based on marginal regression to detect the presence of a significant predictor. Standard inferential methods are known to fail in this setting due to the non-regular limiting behavior of the estimated regression coefficient of the selected predictor; in particular, the limiting distribution is discontinuous at zero as a function of the regression coefficient of the predictor maximally correlated with the outcome. To circumvent this non-regularity, we propose a bootstrap procedure based a local model in order to better reflect small-sample behavior at a root-n scale in the neighborhood of zero. The proposed test is adaptive in the sense that it employs a pre-test to distinguish situations in which a centered percentile bootstrap applies, and otherwise adapts to the local asymptotic behavior of the test statistic in a way that depends continuously on the local parameter. The performance of the approach is evaluated using a simulation study, and applied to an example involving gene expression data.

email: mq2158@columbia.edu

A TWO-DIMENSIONAL APPROACH TO LARGE-SCALE SIMULTANEOUS HYPOTHESIS TESTING USING VORONOI TESSELLATIONS

Daisy L. Phillips*, The Pennsylvania State University
Debashis Ghosh, The Pennsylvania State University

It is increasingly common to see large-scale simultaneous hypothesis tests in which multiple p-values are associated with each test. As a motivating example we consider studies of cell-cycle regulated genes of yeast cells synchronized using different methods. This gives rise to multiple p-values to test for periodicity for each gene. In this paper we propose an approach that accounts for two-dimensional spatial structure of vectors of two p-values (p-vectors) when performing simultaneous hypothesis tests. Our approach uses Voronoi tessellations to incorporate the spatial positioning of p-vectors in the unit square and can be viewed as a bivariate extension of the celebrated Benjamini-Hochberg procedure. We explore various ordering schemes to rank the p-vectors, and use an empirical null approach to control the false discovery rate. We illustrate properties of the new approach using simulations, and apply the approach to Schizosaccharomyces Pombe data.

email: dlp245@psu.edu

A GENERAL MULTISTAGE PROCEDURE FOR K-OUT-OF-N GATEKEEPING

Dong Xi*, Northwestern University
Ajit C. Tamhane, Northwestern University

We offer a generalization of the multistage procedure proposed by Dmitrienko et al. (2008) for parallel gatekeeping to what we refer to as k-out-of-n gatekeeping in which at least k out of n hypotheses ($k = 1,..., n$) in a gatekeeper family must be rejected in order to test the hypotheses in the following family. For $k = 1$ this corresponds to parallel gatekeeping (Dmitrienko et al., 2003) and for $k = n$ to serial gatekeeping (Maurer et al., 1995; Westfall and Krishen, 2001). Besides providing a unified theory of multistage procedures for all $k = 1,..., n$, the paper solves a practical problem that arises in certain situations, e.g., the requirement that efficacy be shown on at least three of the four primary endpoints in trials for rheumatoid arthritis (FDA, 1999).

email: dong.xi@u.northwestern.edu

71. PRESIDENTIAL INVITED ADDRESS

MODELING DATA IN A SCIENTIFIC CONTEXT

Jeremy M. G. Taylor, PhD, University of Michigan

Data are typically collected in a scientific context, with the statistician being part of the team of investigators. The scientific context involves what data are collected and how they are collected, but can also involve scientific knowledge or theories about the underlying mechanisms that give rise to the data. The traditional role of statisticians is to analyze the data and only the data, with an emphasis on using models and methods that make minimal assumptions, that is, "to let the data speak." For confirmatory clinical trials and large epidemiologic studies this may be appropriate. Increasingly, statisticians are involved in laboratory and basic science or other types of studies where the goals are learning, understanding and discovery. Here the data may be multidimensional and complex, and the data analysis may benefit by incorporating scientific knowledge into the analysis models and methods. The assumptions that are incorporated into the models may be mild, such as smoothness or monotonicity; or stronger, such as the existence of a cured group, a coefficient in a regression model being zero, or assumptions about the functional form of a model. In this talk I will discuss the role of models, the bias-variance tradeoff, and distinguish models and methods. I will present case studies from cancer research in which we have incorporated scientific knowledge into the data analysis. Specific examples will include cure models, joint longitudinal-survival models, shrinkage, surrogate endpoints and order-restricted inference.

e-mail: jmgt@umich.edu

72. JABES SHOWCASE

MODELING SPACE-TIME DYNAMICS OF AEROSOLS USING SATELLITE DATA AND ATMOSPHERIC TRANSPORT MODEL OUTPUT

Candace Berrett*, Brigham Young University
Catherine A. Calder, The Ohio State University
Tao Shi, The Ohio State University
Ningchuan Xiao, The Ohio State University
Darla K.. Munroe, The Ohio State University

Kernel-based models for space-time data offer a flexible and descriptive framework for studying atmospheric processes. Nonstationary and anisotropic covariance structures can be readily accommodated by allowing kernel parameters to vary over space and time. In addition, dimension reduction strategies make model fitting computationally feasible for large datasets. Fitting these models to data derived from instruments onboard satellites, which often contain significant amounts of missingness due to cloud cover and retrieval errors, can be difficult. In this presentation, we propose to overcome the challenges of missing satellite-derived data by supplementing an analysis with output from a computer model, which contains valuable information about the space-time dependence structure of the process of interest. We illustrate our approach through a case study of aerosol optical depth across mainland Southeast Asia. We include a crossvalidation study to assess the strengths and weaknesses of our approach.

email: cberrett@stat.byu.edu

UNCERTAINTY ANALYSIS FOR COMPUTATIONALLY EXPENSIVE MODELS

David Ruppert*, Cornell University
Christine A. Shoemaker, Cornell University
Yilun Wang, University of Electronic Science and Technology of China
Yingxing Li, Xiamen University
Nikolay Bliznyuk, University of Florida

MCMC is infeasible for Bayesian calibration and uncertainty analysis of computationally expensive models if one must compute the model at each iteration. To address this problem we introduced SOARS (Statistical and Optimization Analysis using Response Surfaces) methodology. SOARS uses an interpolator as a surrogate, also known as an emulator or meta-model, for the logarithm of the posterior density. To prevent wasteful evaluations of the expensive model, the emulator is only built on a high posterior density region (HPDR), which is located by a global optimization algorithm. The set of points in the HPDR where the expensive model is evaluated is determined sequentially by the GRIMA algorithm. A case study uses an eight-parameter SWAT (Soil and Water Assessment Tool) model where daily stream flows and phosphorus concentrations are modeled for the Town Brook watershed which is part of the New York City water supply.

email: dr24@cornell.edu

IMPROVING CROP MODEL INFERENCE THROUGH BAYESIAN MELDING WITH SPATIALLY-VARYING PARAMETERS

Andrew O. Finley*, Michigan State University
Sudipto Banerjee, University of Minnesota
Bruno Basso, Michigan State University

An objective for applying a Crop Simulation Model (CSM) in precision agriculture is to explain the spatial variability of crop performance. CSMs require inputs related to soil, climate, management, and crop genetic information to simulate crop yield. In practice, however, measuring these inputs at the desired high spatial resolution is prohibitively expensive. We propose a Bayesian modeling framework that melds a CSM with sparse data from a yield monitoring system to deliver location specific posterior predicted distributions of yield and associated unobserved spatially-varying CSM parameter inputs. The proposed Bayesian melding model consists of a systemic component representing output from the physical model and a residual spatial process that compensates for the bias in the physical model. The spatially-varying inputs to the systemic component arise from a multivariate Gaussian process while the residual component is modeled using a univariate Gaussian process. Due to the large number of observed locations in the motivating dataset we seek dimension reduction using low-rank predictive processes to ease the computational burden. The proposed model is illustrated using the Crop Environment Resources Synthesis (CERES)-Wheat CSM and wheat yield data collected in Foggia, Italy.

email: finleya@msu.edu

DEMOGRAPHIC ANALYSIS OF FOREST DYNAMICS USING STOCHASTIC INTEGRAL PROJECTION MODELS

Alan E. Gelfand*, Duke University
Souparno Ghosh, Texas Tech University
James S. Clark, Duke University

Demographic analysis for plant and animal populations is a prominent problem in studying ecological processes, typically using Matrix Projection Models. Integral projection models (IPMs) offer a continuous version of this approach. These models are a class of integro-differential equations which, for demography, we specify a redistribution kernel mechanistically using demographic functions, i.e., parametric models for demographic processes such as survival, growth, and replenishment. With interest in scaling in space, we work with data in the form of point patterns rather than with individual level data (hopeless to scale) yielding intensities (which are easy to scale). Fitting IPMs in our setting is quite challenging and is most feasibly done either by working in the spectral domain or with a pseudo-likelihood, in conjunction with Laplace approximation. We illustrate with an investigation of forest dynamics using data from Duke Forest as well as a U.S.national survey called the Forest Inventory Analysis.

email: alan@stat.duke.edu

73. STATISTICAL CHALLENGES IN LARGE-SCALE GENETIC STUDIES OF COMPLEX DISEASES

GENE-GENE INTERACTION ANALYSIS FOR NEXT-GENERATION SEQUENCING

Momiao Xiong*, University of Texas School of Public Health
Yun Zhu, Tulane University
Futao Zhang, University of Texas School of Public Health

Critical barriers in interaction analysis for rare variants is that most traditional statistical methods for testing gene-gene interaction were originally designed for testing the interaction for common variants and are difficult to be applied to rare variants due to their low power. The great challenges for successful detection of interactions with next-generation sequencing data are (1) lack of deep understanding measure of interaction and statistics with high power to detect interaction, (2) lack of concepts, methods and tools for detection of interactions for rare variants, (3) severe multiple testing problems, and (4) heavy computations. To meet this challenge, we take a genome region or a gene as a basic unit of interaction analysis and use high dimensional data reduction techniques to develop a novel statistic for collectively test interaction between all possible pairs of SNPs within two genome regions or genes. By large-scale simulations, we demonstrate that the proposed new statistic has the correct type 1 error rates and much higher power than the existing methods to detect gene-gene. To further evaluate its performance, the developed statistic is applied to the lip metabolism trait exome sequence data from the NHLBI's Exome Sequencing Project (ESP) and whole genome-sequence data.

email: momiao.xiong@uth.tmc.edu

ASSOCIATION MAPPING OF RARE VARIANTS IN SAMPLES WITH RELATED INDIVIDUALS

Duo Jiang, University of Chicago
Mary Sara McPeek*, University of Chicago

One fundamental problem of interest is to identify genetic variants that contribute to observed variation in human complex traits. With the increasing availability of high-throughput sequencing data, there is the possibility of identifying rare variants that influence a trait, but there

may be low power to detect association with any individual rare variant. By combining information across a group of rare variants in a gene or pathway, it is possible to increase power. Many genetic studies contain data on related individuals, and such studies can be particularly helpful for identifying and validating rare variants. We describe statistical methods for mapping rare variants in samples with completely general combinations of families and unrelated individuals, including large complex pedigrees.

email: msmcpeek@gmail.com

HIDDEN HERITABILITY AND RISK PREDICTION BASED ON GENOME-WIDE ASSOCIATION STUDIES

Nilanjan Chatterjee*, National Cancer Institute, National Institutes of Health
JuHyun Park, National Cancer Institute, National Institutes of Health

We report a new model to project the predictive performance of polygenic models based on the number and distribution of effect sizes for the underlying susceptibility alleles, the size of the training dataset and the balance of true and false positives among loci selected in the models. Using effect-size distributions derived from discoveries from the largest genome-wide association studies and estimates of hidden heritability, we assess predictive ability of common Single Nucleotide Polymorphisms (SNP) for ten complex traits. We project that while 45% of the total variance of adult height has been attributed to common SNPs, a model built based on one million people may only explain 33.4% of variance of the trait in an independent sample. Models built based on current GWAS can identify 3.0%, 1.1%, and 7.0%, of the populations who are at two-fold or higher than average risk for Type 2 diabetes, coronary artery disease and prostate cancer, respectively. By tripling the sample size in the future, the corresponding percentages could be elevated to 18.8%, 6.1%, and 12.2%, respectively. The predictive utility of future polygenic models will depend not only on heritability, but also on achievable sample sizes, effect-size distribution and information on other risk-factors, including family history.

email: chattern@mail.nih.gov

ON A CLASS OF FAMILY-BASED ASSOCIATION TESTS FOR SEQUENCE DATA, AND COMPARISONS WITH POPULATION-BASED ASSOCIATION TESTS

Iuliana Ionita-Laza*, Columbia University
Seunggeun Lee, Harvard University
Vlad Makarov, Mount Sinai School of Medicine
Joseph Buxbaum, Mount Sinai School of Medicine
Xihong Lin, Harvard University

Recent advances in high-throughput sequencing technologies make it increasingly more efficient to sequence large cohorts for many complex traits. We discuss here a class of sequence-based association tests for family-based designs that corresponds naturally to previously proposed population-based tests, including the classical burden and variance-component tests. This framework allows for a direct comparison between the power of sequence-based association tests with family- vs. population-based designs. We show that, for dichotomous traits using family-based controls results in similar power levels as the population-based design (although at an increased sequencing cost for the family-based design), while for continuous traits (in random samples, no ascertainment) the population-based design can be substantially more powerful. A possible disadvantage of population-based designs is that they can lead to increased false-positive rates in the presence of population stratification, while the family-based designs are robust to population stratification. We show also an application to a small exome-sequencing family-based study on autism spectrum disorders. The tests are implemented in publicly available software.

email: ii2135@columbia.edu

74. ANALYSIS OF HIGH-DIMENSIONAL DATA

STOCHASTIC OPTIMIZATION FOR SPARSE HIGH-DIMENSIONAL STATISTICS: SIMPLE ALGORITHMS WITH OPTIMAL CONVERGENCE RATES

Alekh Agarwal, Microsoft Research
Sahand Negahban, Massachusetts Institute of Technology
Martin J. Wainwright*, University of California, Berkeley

Many effective estimators for high-dimensional statistical problems, such as sparse regression or low-rank matrix estimation, are based on solving large-scale convex optimization problems. The high dimensional nature makes simple methods, such as the on-line methods of stochastic optimization, particularly attractive. Most existing methods either obtain a slow $\order(1/\sqrt{T})$ convergence rate with a mild dimension dependence, or a faster $\order(1/T)$ convergence with a poor scaling in dimension. Building on recent work of Juditsky and Nesterov, we develop an algorithm that enjoys $\order(1/T)$ convergence, while maintaining a mild dimension dependence. For the special case of sparse linear regression, this results in a one-pass stochastic algorithm with convergence rate that scales optimally with the number of iterations T, number of dimensions d and the sparsity level s. We also complement our theory with numerical simulations on large-scale problems. Based on joint work with Alekh Agarwal and Sahand Negahban http://arxiv.org/abs/1207.4421.

email: wainwrig@stat.berkeley.edu

MINIMAX AND ADAPTIVE ESTIMATION OF COVARIANCE OPERATOR FOR RANDOM VARIABLES OBSERVED ON A LATTICE GRAPH

Tony Cai, University of Pennsylvania
Ming Yuan*, Georgia Tech

Covariance structure plays an important role in high dimensional statistical inference. In a range of applications including imaging analysis and fMRI studies, random variables are observed on a lattice graph. In such a setting it is important to account for the lattice structure when estimating the covariance operator. We consider here both minimax and adaptive estimation of the covariance operator over collections of polynomially decaying and exponentially decaying parameter spaces.

email: myuan@isye.gatech.edu

STRONG ORACLE PROPERTY OF FOLDED CONCAVE PENALIZED ESTIMATION

Jianqing Fan, Princeton University
Lingzhou Xue, Princeton University
Hui Zou*, University of Minnesota

Folded concave penalization methods (Fan and Li, 2001) have been shown to enjoy the strong oracle property for high-dimensional sparse estimation. However, a folded concave penalization problem usually has multiple local solutions and the oracle property is established only for one of the unknown local solutions. A challenging fundamental issue still remains that it is not clear whether the local optimal solution computed by a given optimization algorithm possesses those nice theoretical properties. To close this important theoretical gap in over a decade, we provide a unified theory to show explicitly how to obtain the oracle solution using the local linear approximation algorithm. For a folded concave penalized estimation problem, we show that as long as the problem is localizable and the oracle estimator is well behaved, we can obtain the oracle estimator by using the one-step local linear approximation. In addition, once the oracle estimator is obtained, the local linear approximation algorithm converges, namely produces the same estimator in the next iteration. The general theory is demonstrated by using three classical sparse estimation problems, i.e. the sparse linear regression, the sparse logistic regression and the sparse precision matrix estimation.

email: zouxx019@umn.edu

SIMULTANEOUS AND SEQUENTIAL INFERENCE OF PATTERN RECOGNITION

Wenguang Sun*, University of Southern California

The accurate and reliable recovery of sparse signals in massive and complex data has been a fundamental question in many scientific fields. The discovery process

usually involves an extensive search among a large number of hypotheses to separate signals of interest and also recognize their patterns. The situation can be described as finding needles of various shapes in a haystack. Despite the enormous progress on methodological work in data screening, pattern recognition and related fields, there have been little theoretical studies on the issues of optimality and error control in situations where a large number of decisions are made sequentially and simultaneously. We develop a compound decision theoretic framework and propose a new loss matrix approach to generalize the current multiple testing framework for error control in pattern recognition, by allowing more than two states of nature, sequential decision-making and new concepts of false positive rates in large-scale simultaneous inference.

email: wenguans@marshall.usc.edu

75. STATISTICAL BODY LANGUAGE: ANALYTICAL METHODS FOR WEARABLE COMPUTING

A NOVEL METHOD TO ESTIMATE FREE-LIVING ENERGY EXPENDITURE FROM AN ACCELEROMETER

John W. Staudenmayer*, University of Massachusetts, Amherst
Kate Lyden, University of Massachusetts, Amherst

The purpose of this paper is to develop and validate a novel method for estimating energy expenditure in free-living people. The method uses an accelerometer, a device that measures and records the quantity and intensity of movement, and an algorithm to estimate energy expenditure from the resulting accelerometer signals. The proposed method is a two-step process. In the first step, we use simple characteristics of the acceleration signal to identify where bouts activity and inactivity start and stop. In the second step, we estimate energy expenditure (METs) for each bout using methods that were previously developed and validated in the laboratory on several hundred people. We compare the proposed algorithm, which we call the "sojourn algorithm," to existing methods using data from a group of individuals who were each directly observed over the course of two 10-hour days. The sojourn algorithm is more accurate and precise than existing methods. The new algorithm is specifically designed for use in free-living environments where behavior is not planned and does not occur in intervals of known duration. It also has the potential to provide more detailed information about information about the duration and frequency of bouts of activity and inactivity.

email: jstauden@math.umass.edu

HEART-TO-HEART DIARY OF PHYSICAL ACTIVITY

Vadim Zipunnikov*, Johns Hopkins Bloomberg School of Public Health
Jennifer Schrack, Johns Hopkins Bloomberg School of Public Health
Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health
Jeff Goldsmith, Columbia University
Luigi Ferrucci, National Institute on Aging, National Institutes of Health

Physical activity energy expenditure is a modifiable risk factor for multiple chronic diseases. Accurate measurement of free-living energy expenditure is vital to understanding and quantifying changes in energy metabolism and their effects on activity, disability, and disease in population-based studies. Actiheart is a novel device that monitors minute-by-minute activity counts and heart rate and uses both to estimate energy expenditure. We will first illustrate key statistical challenges with data collected on 794 subjects wearing Actiheart for one week as a part of the Baltimore Longitudinal Study of Aging. We will then describe the methods to address those challenges and parsimoniously model a complex interdependence between activity and heart rate. At the end, we will show how our models can be used to construct a population-based dynamic diary that describes and quantifies the most common daily patterns of activity.

email: vzipunni@jhsph.edu

A NEW ACCELEROMETER WEAR AND NONWEAR TIME CLASSIFICATION ALGORITHM

Leena Choi*, Vanderbilt University School of Medicine
Suzanne C. Ward, Vanderbilt University School of Medicine
John F. Schnelle, Vanderbilt University School of Medicine
Maciej S. Buchowski, Vanderbilt University School of Medicine

The use of accelerometers for measuring physical activity (PA) in intervention and population-based studies is becoming a standard methodology for the objective measurement of sedentary and active behaviors. Data collected by accelerometers such as ActiGraph in a natural free-living environment can be divided into wear and nonwear time intervals. Since these accelerometer data are very large, it is not feasible to manually classify nonwear and wear time intervals without using an automated algorithm. Thus, a vital step in PA measurement is the classification of daily time into accelerometer wear and nonwear intervals using its recordings (counts) and an accelerometer-specific algorithm. Typically, an automated algorithm uses monitor-specific criteria to detect and to eliminate the nonwear time intervals, during which no activity is detected. The most commonly used automated algorithm for ActiGraph data is based on the criteria proposed by Troiano (2007), which has been used in several population based studies, including the National Health and Nutrition Examination Survey (NHANES). We evaluated and improved this algorithm for both uniaxial and triaxial ActiGraph data. The improved algorithm classified wear and nonwear intervals more accurately, and may lead to more accurate estimation of time spent in sedentary and active behaviors.

email: leena.choi@vanderbilt.edu

QUANTIFYING PHYSICAL ACTIVITY USING ACCELEROMETERS

Julia Kozlitina*, University of Texas Southwestern Medical Center
William R. Schucany, Southern Methodist University

Accelerometers have become a widely used tool for the objective assessment of physical activity (PA) in large epidemiological and surveillance studies. Despite their widespread use, many questions remain regarding the processing and interpretation of accelerometer output in order to derive accurate measures of PA. Traditionally, accelerometer data have been summarized as time spent in different PA intensities, using established cut-points to classify activity into intensity levels. To reflect the accumulation of activity that may be meaningful for energy balance, the time spent above various intensity thresholds is further summarized into continuous 10-min bouts. This suggests that local averaging may be helpful in identifying intervals of PA corresponding to different intensity levels. We use different smoothing techniques when extracting PA bout information and examine their impact on the resulting outcome variables. We illustrate the analysis using objectively measured activity data from a large population-based study.

email: Julia.Kozlitina@UTSouthwestern.edu

76. BIOMARKER UTILITY IN CLINICAL TRIALS

DESIGN AND ANALYSIS OF BIOMARKER THRESHOLD STUDIES IN RANDOMIZED CLINICAL TRIALS

Glen Laird*, Bristol-Myers Squibb
Yafeng Zhang, University of California, Los Angeles

As the importance of individualized medicine grows, the use of biomarkers to guide population selection is becoming more common. Testing for significant effects in a biomarker-defined subpopulation can improve statistical power and permit focused treatment on patients most likely to benefit. A key biomarker may be measured on a continuous scale, but a binary classification of patients is convenient for treatment purposes. Towards this goal, determination of a biomarker threshold level may be be

necessary in an environment where the efficacy of both the experimental and control treatments is correlated with the biomarker. We review the task of classifying patients into the subpopulation and compare operating characteristics for some implementation differences.

email: glen.laird@bms.com

BIOMARKER UTILITY IN DRUG DEVELOPMENT PROGRAMS

Christopher L. Leptak*, U.S. Food and Drug Administration

As drug development tools, biomarkers hold promise in aiding drug development programs by introducing innovative approaches to challenges currently facing the industry. Broadly defined, biomarkers have utility throughout the drug development process and can significantly contribute to the overall benefit-risk assessment. Although biomarker acceptance has historically occurred over extended periods of time through the accumulation of scientific knowledge and experience, current proactive approaches strive to make biomarker identification more efficient through a more focused, data-driven process. The presentation will highlight the roles biomarkers may play in clinical trial design, pathways that can lead to effective biomarker development, and examples of emerging best practices.

email: christopher.leptak@fda.hhs.gov

ANALYSIS OF INTERACTIONS FOR ASSESSING HETEROGENEITY OF TREATMENT EFFECT IN A CLINICAL TRIAL

Stephanie A. Kovalchik*, National Cancer Institute, National Institutes of Health
Carlos O. Weiss, Johns Hopkins University
Ravi Varadhan, Johns Hopkins University

A subgroup analysis is a comparison of treatment effect between subsets of patients in a clinical trial. Though they are recognized to have high false-positive and fase-negative rates, clinical trials frequently report multiple subgroup analyses. When treatment responsiveness depends on multiple patient characteristics, as current reporting practice suggests, simultaneous assessment of effect modification could have greater power than univariate approaches. We, therefore, investigated the operational characteristics of approaches to quantify joint treatment-covariate interactions, denoted analysis of interactions (ANOINT). In addition to conventional one-by-one testing, we considered an unstructured joint interaction model, a proportional interactions model, and sequential procedures combining these approaches for regression models in the generalized linear or proportional hazards families. Simulation studies were used to evaluate the performance of the methods under different trial and model selection scenarios. We present recommendations for the assessment of heterogeneity of treatment effect in clinical trials based on our findings. The ANOINT methodology is illustrated with an analysis of the Studies of Left Ventricular Dysfunction Trial. The presented methods can be implemented with our open-source R package anoint.

email: kovalchiksa@mail.nih.gov

BIOMARKER SELECTION AND ESTIMATION WITH HETEROGENEOUS POPULATION

Shuangge Ma*, Yale University

Heterogeneity inevitably exists across subpopulations in clinical trials and observational studies. If such heterogeneity is not properly accounted for, the selection of biomarkers and estimation of their effects can be biased. Two models are proposed to describe biomarker selection when heterogeneity exists. Under the homogeneity model, the importance (relevance) of a biomarker is consistent across multiple subpopulations, however, its effect may vary. In contrast, under the heterogeneity model, the importance of a biomarker may also vary across multiple subpopulations, suggesting that such a biomarker may be only applicable to certain subpopulations. Multiple regularization methods are proposed, tailoring biomarker selection under different scenarios. Simulation study with moderate to high dimensional data demonstrates reasonable performance of the proposed selection methods. Analysis of cancer survival studies shows that accounting for heterogeneity is necessary, and the proposed methods can identify important biomarkers missed by the existing studies.

email: shuangge.ma@yale.edu

77. NOVEL APPROACHES FOR MODELING VARIANCE IN LONGITUDINAL STUDIES

JOINT MODELING OF LONGITUDINAL HEALTH PREDICTORS AND CROSS-SECTIONAL HEALTH OUTCOMES VIA MEAN AND VARIANCE TRAJECTORIES

Bei Jiang, University of Michigan
Michael Elliott*, University of Michigan
Naisyin Wang, University of Michigan
Mary D. Sammel, University of Pennsylvania Perelman School of Medicine

Growth Mixture Models (GMMs) are used to model heterogeneity in longitudinal trajectories. GMMs assume that each subject's growth curve, characterized by random coefficients in mixed effects models, belongs to an underlying latent cluster with a cluster- specific mean profile. Within-subject variability is typically treated as a nuisance and assumed to be non-differential. Elliott (2007) extended the idea of modeling random effects as finite mixtures as in GMMs into the variance structure setting, where underlying 'clusters' of within-subject variabilities were related to the health outcome of interest while the subject-specific trajectories were treated entirely as nuisance and modeled by penalized smoothing splines. We extend these ideas by allowing 'heterogeneities' (i.e., clusters) in both the growth curves and within-subject variabilities and develop a method that simultaneously examines the association between the underlying mean growth profile and the variance clusters with a cross-sectional binary health outcome. We consider an applications to predict onset of senility in a population sample of older adults using memory test scores and to predict severe hot flashes using the hormone levels collected over time for women in menopausal transition.

email: mrelliot@umich.edu

DETANGLING THE EFFECT BETWEEN RATE OF CHANGE AND WITHIN-SUBJECT VARIABILITY IN LONGITUDINAL RISK FACTORS AND ASSOCIATIONS WITH A BINARY HEALTH OUTCOME

Mary D. Sammel*, University of Pennsylvania Perelman School of Medicine

To evaluate the effect of longitudinally measured risk factors on the subsequent development of disease, it is often necessary to calculate summary measures of these factors to capture features of the risk profile. These methods typically consider correlations among repeated measurements on subjects as nuisance parameters. Using an example of hormone profile changes in the menopausal transition, we demonstrate that residual variability in subject measures, in addition to risk factor profiles, may also be important in predicting future health outcomes of interest. We explore joint models allowing us to structure within- and between-subject variability from longitudinal studies, and combine variance structures with mean structures such as mean longitudinal profiles to better understand the relationship between longitudinally measured risk factors and health outcomes. In the context of a 14 year longitudinal cohort study of ovarian aging among women approaching the menopause we hypothesized that fluctuations in hormone levels, rather than absolute levels, predicted menopausal symptoms. Increased efficiency in estimating associations of interest are demonstrated over a simplified 2 stage estimation approach.

email: msammel@upenn.edu

A LOCATION SCALE ITEM RESPONSE THEORY (IRT) MODEL FOR ANALYSIS OF ORDINAL QUESTIONNAIRE DATA

Donald Hedeker*, University of Illinois at Chicago
Robin J. Mermelstein, University of Illinois at Chicago

Questionnaires are commonly used in studies of health to measure severity of illness, for example, and the items are often scored on an ordinal scale. For such questionnaires, item response theory (IRT) models provide a useful approach for obtaining summary scores for subjects (the model's random subject effect) and characteristics of the items (item difficulty and discrimination). We describe an extended IRT model that allows the items to exhibit different within-subject (WS) variance, and also include a subject-level random effect to the WS variance specification. This permits subjects to be characterized in terms of their mean level, or location, and their variability, or scale. We illustrate application of this location scale IRT model using data from the Nicotine Dependence Syndrome Scale (NDSS) assessed in an adolescent smoking study. We show that there is an interaction between a subject's mean and scale in predicting future smoking level, such that, for low-level smokers, increased scale is associated with subsequent higher smoking levels. The proposed location scale IRT model has useful applications in research where questionnaires are often rated on an ordinal scale, and there is interest in characterizing subjects in terms of both their mean and variance.

email: hedeker@uic.edu

BAYESIAN MIXED-EFFECTS LOCATION SCALE MODELS FOR THE ANALYSIS OF OBJECTIVELY MEASURED PHYSICAL ACTIVITY DATA FROM A LIFESTYLE INTERVENTION TRIAL

Juned Siddique*, Northwestern University Feinberg School of Medicine
Donald Hedeker, University of Illinois at Chicago

Objective measurement of physical activity using wearable accelerometers is now used in large-scale epidemiological studies and clinical trials of lifestyle and exercise interventions. These devices measure the frequency, intensity, and duration of physical activity at the momentary level, often using measurement epochs of 1 minute or less yielding a large number of observations per subject. In this talk, we describe a Bayesian mixed-effects location scale model for the analyses of physical activity as measured by accelerometer using data from a randomized lifestyle intervention trial of 204 men and women. We model both the mean and variance over time as a function of subject-specific random effects and time varying covariates. Our model allows us to measure how covariates can influence both the mean and the variance of physical activity over time. It also allows us to measure whether changes in variability over the course of the study are predictive of treatment relapse.

email: siddique@northwestern.edu

78. EVIDENCE SYNTHESIS FOR ASSESSING BENEFIT AND RISK

SYSTEMATIC REVIEWS IN COMPARATIVE EFFECTIVENESS RESEARCH

Sally C. Morton*, University of Pittsburgh

Systematic reviews in comparative effectiveness research (CER) are essential to identify gaps in evidence for making decisions that matter to stakeholders, most notably patients. In this talk, I will discuss unique issues that may arise in a CER systematic review. For example, such systematic reviews may include observational data in order to assess both benefits and harms. Methodological techniques such as network meta-analysis may be required to compare treatments that have the potential to be best practice. The risk of bias for individual studies must be assessed to estimate effectiveness in real-world settings. Finally, the strength of the body of evidence will need to be determined to inform decision-making. Current guidance from the Agency for Healthcare Research and Quality (AHRQ); the Cochrane Collaboration; and the Institute of Medicine (IOM) will be included.

email: scmorton@pitt.edu

BAYESIAN INDIRECT AND MIXED TREATMENT COMPARISONS ACROSS LONGITUDINAL TIME POINTS

Haoda Fu*, Eli Lilly and Company
Ying Ding, Eli Lilly and Company

Meta-analysis has become an acceptable and powerful tool for pooling quantitative results from multiple studies addressing the same question. It estimates the effect difference between two treatments when they have been compared head-to-head. However, limitations occur when there are more than two treatments of interest and some of them have not been compared in the same study. Indirect and mixed treatment modeling extends meta-analysis methods to enable data from different treatments and trials to be synthesized, without requiring head-to-head comparisons among all treatments; thus, allowing different treatments can be compared. Traditional indirect and mixed treatment comparison methods consider a single endpoint for each trial. We extend the current methods and propose a Bayesian indirect and mixed treatment comparison longitudinal model. That incorporates multiple time points and allows indirect comparisons of treatment effects across different longitudinal studies. Simulation studies were performed which demonstrate that the proposed method performs well and yields better estimations compared to other single time point meta-analysis methods. We apply our method to a set of studies from patients with type 2 diabetes.

email: fuhaoda@gmail.com

ADAPTIVE TRIAL DESIGN IN THE PRESENCE OF HISTORICAL CONTROLS

Brian P. Hobbs*, University of Texas MD Anderson Cancer Center
Bradley P. Carlin, University of Minnesota
Daniel J. Sargent, Mayo Clinic

Prospective trial design often occurs in the presence of 'acceptable' (Pocock, 1976) historical control data. Typically this data is only utilized for treatment comparison in a posteriori, retrospective analysis. We propose an adaptive trial design in the context of an actual randomized controlled colorectal cancer trial. The proposed trial implements an adaptive randomization procedure for allocating patients aimed at balancing total information (concurrent and historical) among the study arms. This is accomplished by assigning more patients to receive the novel therapy in the absence of strong evidence for heterogeneity among the concurrent and historical controls. Allocation probabilities adapt as a function of the effective historical sample size (EHSS) characterizing relative informativeness defined in the context of a piecewise exponential model for evaluating time to disease progression. A Bayesian hierarchical model is used to assess historical and concurrent heterogeneity at interim analyses and to borrow strength from the historical data in the final analysis. Using the proposed hierarchical model to borrow strength from the historical data, after balancing total information with the adaptive randomization procedure, provides preposterior admissible estimators of the novel treatment effect with desirable bias-variance trade-offs.

email: bphobbs@mdanderson.org

INCORPORATING EXTERNAL INFORMATION TO ASSESS ROBUSTNESS OF COMPARATIVE EFFECTIVENESS ESTIMATES TO UNOBSERVED CONFOUNDING

Mary Beth Landrum*, Harvard Medical School
Alfa Alsane, Harvard Medical School

Successful reform of the health care delivery system relies on improved information about the effectiveness of therapies in real world practice. While comparative effectiveness research often relies on synthesis of evidence from randomized clinical trials to infer effectiveness of

therapies, many rely on the analysis of observational data sources where patients are not randomized to treatment. Comparative effectiveness studies based on observational data are reliant on a set of assumptions that often cannot be tested empirically. Sensitivity analyses attempt to examine the robustness of estimates to plausible violations of these key assumptions. A variety of different methods to assess the robustness of estimates have been proposed and applied in the context of propensity score analyses. Recent work has also proposed the use of sensitivity analyses in IV analyses. These approaches include simple calculations of how estimates change under certain violations of assumptions and Bayesian approaches that make parameters governing key assumptions part of the model, thereby allowing uncertainty about violations of assumptions to be incorporated in statistical inferences. In this talk we compare the performance of various approaches in the context of an analysis of the comparative effectiveness of cancer therapies in elderly populations typically underrepresented in cancer RCTs.

email: landrum@hcp.med.harvard.edu

79. MODEL SELECTION AND ANALYSIS IN GWAS STUDIES

A GENETIC RISK PREDICTION METHOD BASED ON SVM

Qianchuan He*, Fred Hutchinson Cancer Research Center
Helen Zhang, North Carolina State University
Dan-Yu Lin, University of North Carolina, Chapel Hill

Genetic risk prediction plays an important role in the era of personalized medicine. The task of genetic risk prediction is highly challenging, as many complex human diseases are contributed by a large number of genetic variants, many of which with relatively small effects. This task is further complicated by high correlations among SNPs, low penetrance of the causal variants, and unknown genetic models of the underlying risk loci. We propose a new method that is based on the Sure Independence Screening and the SCAD-SVM (smoothly clipped absolute deviation-support vector machine). This method is effective in reducing the originally large number of predictors into a relatively small set of predictors, and appears to be robust to different underlying genetic models. Simulation studies and real data analysis show that the proposed method has some advantages over existing methods.

email: qhe@fhcrc.org

TESTING GENETIC ASSOCIATION WITH BINARY AND QUANTITATIVE TRAITS USING A PROPORTIONAL ODDS MODEL

Gang Zheng, National Heart, Lung and Blood Institute, National Institutes of Health
Ruihua Xu*, George Washington University
Neal Jeffries, National Heart, Lung and Blood Institute, National Institutes of Health
Ryo Yamada, Kyoto University
Colin O. Wu, National Heart, Lung and Blood Institute, National Institutes of Health

When a genetic marker is associated with binary and quantitative traits individually, testing a joint association of the marker with both traits is more powerful than testing a single trait if the traits are correlated. In this paper, we propose a simple approach to combine two mixed types of traits into a single ordinal outcome and apply a proportional odds model to detect pleiotropic association. We demonstrate how this proposed test relates to the associations with individual traits. Simulation results are presented to compare the proposed method with existing ones. The results are applied to a genome-wide association study of rheumotoid arthritis with anticyclic citrullinated peptide.

email: rxu@gwmail.gwu.edu

A NOVEL PATHWAY-BASED ASSOCIATION ANALYSIS WITH APPLICATION TO TYPE 2 DIABETES

Tao He*, Michigan State University
Yuehua Cui, Michigan State University

Single-marker-based tests in genome-wide association studies (GWAS) have been very successful in identifying thousands of genetic variants associated with hundreds of complex diseases. However, these identified variants only explain a small fraction of inheritable variability, suggesting that other factors, such as multilevel genetic variations and gene×environment interactions may contribute to disease susceptibility more than we anticipated. In this work, we propose to combine genetic signals at different levels, such as in gene- and pathway-level to form integrated signals aimed to identify major players that function in a coordinated manner. The integrated analysis provides novel insight into disease etiology while the signals could be easily missed by single variants analysis. We demonstrate the merits of our method through simulation studies. Finally, we applied our approach to a genome-wide association study of type 2 diabetes (T2D) with male and female data analyzed separately. Novel sex-specific genes and pathways were identified to increase the risk of T2D.

email: hetao@msu.edu

TEST FOR INTERACTIONS BETWEEN A GENETIC MARKER SET AND ENVIRONMENT IN GENERALIZED LINEAR MODELS

Xinyi (Cindy) Lin*, Harvard School of Public Health
Seunggeun Lee, Harvard School of Public Health
David C. Christiani, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

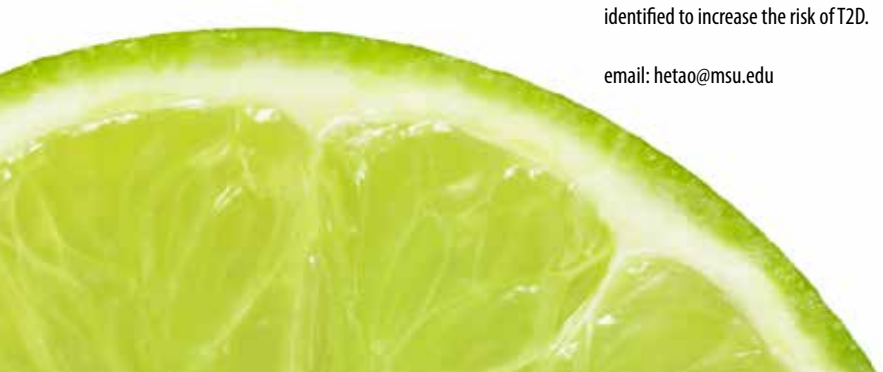
We consider testing for interactions between a genetic marker set and an environmental variable in genome wide association studies. A common practice in studying gene-environment (GE) interactions is to analyze one SNP at a time in a classical GE interaction model and adjust for multiple comparisons across the genome. It is of significant current interest to analyze SNPs in a set simultaneously. In this paper, we first show that if the main effects of multiple SNPs in a set are associated with a disease/trait, the classical single SNP GE interaction analysis is generally biased. We derive the asymptotic bias and study the conditions under which the classical single SNP GE interaction analysis is unbiased. We further show that, the simple minimum p-value based SNP-set GE analysis, which calculates the minimum of the GE interaction p-values by fitting individual GE interaction models for each SNP in the SNP-set separately is often biased and has an inflated Type 1 error rate. To overcome these difficulties, we propose a computationally efficient and powerful gene-environment set association test (GESAT) in generalized linear models. We evaluate the performance of GESAT using simulation studies, and apply GESAT to data from the Harvard lung cancer genetic study to investigate GE interactions between the SNPs in the 15q24-25.1 region and smoking on lung cancer risk.

email: xinyilin@fas.harvard.edu

LEVERAGING LOCAL IBD INCREASES THE POWER OF CASE/CONTROL GWAS WITH RELATED INDIVIDUALS

Joshua N. Sampson*, National Cancer Institute, National Institutes of Health
Bill Wheeler, Information Management Services
Peng Li, National Cancer Institute, National Institutes of Health
Jianxin Shi, National Cancer Institute, National Institutes of Health

Genome-Wide Association Studies (GWAS) of binary traits can include related individuals from known pedigrees. Until now, standard GWAS analyses have focused on the individual. For each individual, a study records their genotype and disease status. We introduce a GWAS analysis that takes a different perspective and focuses on the founder chromosomes within each family. For each founder chromosome, we effectively identify its allele (at a given SNP) and the proportion of individuals carrying that chromosome who are affected. This step is made possible by recent advances in Identity-By-Descent



(IBD) mapping and haplotyping. We then suggest a chromosome-based Quasi Likelihood Score (cQLS) score statistic that, at its simplest, measures the correlation between the alleles and proportion of individuals affected, and show that tests based on this statistic can increase power.

email: joshua.sampson@nih.gov

A NOVEL METHOD TO EVALUATE THE NONLINEAR RESPONSE OF MULTIPLE VARIANTS TO ENVIRONMENTAL STIMULI

Yuehua Cui, Michigan State University
Cen Wu*, Michigan State University

A variety of system-based (such as gene and/or pathway based) analysis in genome-wide association studies (GWAS) have demonstrated promising prospects in targeting susceptibility loci correlated with complex diseases. The advantages of these methods are especially prominent when multiple variants function jointly in a complicated manner. In this work, we extend our previous method on the dissection of non-linear genetic penetrance of a single variant in gene-environment (G×E) interactions using varying coefficient (VC) model to gene based multiple variant analysis, given that these variants are mediated by a common environment factor. Evaluating the nonlinear mechanism of genetic variants in such a group manner via the additive VC model has particular power to help us elucidate the complicated genetic machinery of G×E interaction that increase disease risk. Assessing overall genetic G×E interaction and nonlinear interaction effects is done via a simultaneous variable selection approach corresponding to selecting zero, constant and varying coefficients, respectively. The merit of the proposed method is evaluated through extensive simulation studies and real data analysis.

email: wucen@stt.msu.edu

WEIGHTED COMPOSITE LIKELIHOOD FOR ANALYSIS OF MULTIPLE SECONDARY PHENOTYPES IN GENETIC ASSOCIATION STUDIES

Elizabeth D. Schifano*, University of Connecticut
Tamar Sofer, Harvard School of Public Health
David C. Christiani, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

There is increasing interest in the joint analysis of multiple phenotypes in genome-wide association studies (GWAS), especially for analysis of multiple secondary phenotypes in case-control studies. By taking advantage of correlation, both across phenotypes and across Single Nucleotide Polymorphisms (SNPs), one could potentially gain statistical power. We propose novel statistical testing and variable selection procedures based on composite likelihood theory to identify SNP sets (e.g., SNPs grouped within a gene), as well as individual SNPs, associated with multiple phenotypes. As such, both procedures allow for adjustment of covariate effects and are robust to misspecification of the true correlation across outcomes. For multiple secondary phenotypes, we use a weighted composite likelihood to account for case-control ascertainment in testing and variable selection, and additionally propose a weighted Bayesian Information Criterion for tuning parameter selection. We demonstrate the effectiveness of both procedures through theoretical and empirical analysis, as well as in application to investigate SNP associations with smoking behavior measured using multiple secondary smoking phenotypes in a lung cancer case-control GWAS.

email: elizabeth.schifano@uconn.edu

80. ADAPTIVE DESIGN AND RANDOMIZATION

INFERENCES FOR COVARIATE-ADJUSTED RESPONSE-ADAPTIVE DESIGNS

Hongjian Zhu*, University of Texas School of Public Health, Houston
Feifang Hu, University of Virginia

Covariate-adjusted response-adaptive (CARA) designs, which sequentially allocate the next patient to treatments in a clinical trial based on the full history of previous treatment assignments, responses, covariates and the current covariate profile, have been shown to be advantageous over traditional designs in terms of both ethics and efficiency. But its subjects allocating manner makes the structure of allocated response sequences very complicated. Accordingly, the validation of statistical inferences is usually challenging, and few theoretical results have been obtained. In this paper, we try to solve fundamental problems for general CARA designs. First, we obtained the conditional independence and distribution of the allocated response sequences, which is the basis of further theoretical investigations. More importantly, we proposed a framework for proposing new CARA designs with unified asymptotic results for statistical inferences. Besides, with the help of an explicit conclusion about the relationship of the allocated response sequences, we also improved the CARA design proposed by Zhang et al. (2007), allowing their design to be applied to many more common models. Understanding of the above fundamentals of CARA designs will potentially promote its development and application.

email: hongjian.zhu@uth.tmc.edu

STATISTICAL INFERENCE OF COVARIATE-ADAPTIVE RANDOMIZED CLINICAL TRIALS

Wei Ma*, University of Virginia
Feifang Hu, University of Virginia

It is well known that covariates often play important roles in clinical trials. Covariate-adaptive designs are usually implemented to balance important covariates in clinical studies. However, to analysis covariate-adaptive randomized clinical trials, the properties of conventional statistical inference methods are usually unknown. Some simulated studies show that testing hypotheses are too conservative. In this talk, we study the theoretical properties of hypothesis testing based on linear model under covariate-adaptive designs. Theoretically we prove that the testing hypotheses are usually conservative in terms of small Type I errors under the null hypothesis under a large class of covariate-adaptive designs. The class includes the two popular covariate-adaptive randomization procedures: Pocock and Simon's marginal procedure (Pocock and Simon, 1975) and stratified permuted block design. Numerical studies are performed to assess Type I errors and power comparison. Some methods to adjust Type I errors are also studied and recommended.

email: wm5yh@virginia.edu

INFORMATION-BASED SAMPLE SIZE RE-ESTIMATION IN GROUP SEQUENTIAL DESIGN FOR LONGITUDINAL TRIALS

Jing Zhou*, University of North Carolina, Chapel Hill
Adeniyi Adewale, Merck
Yue Shentu, Merck
Jiajun Liu, Merck
Keaven Anderson, Merck

Group sequential design has become more popular in clinical trials since it allows for trials to stop early for futility or efficacy to save time and resources. However, this approach is less well-known for longitudinal analysis. We have observed repeated cases of studies with longitudinal data where there is an interest in early stopping for a lack of treatment effect or in adapting sample size to correct for inappropriate variance assumptions. We propose an information-based group sequential design as a method to deal with both of these issues. Updating the sample size at each interim analysis will allow us to maintain the target power and control the type-I error rate. We will illustrate our strategy by real data analysis examples and simulations and compare the results with those obtained using fixed design and group-sequential design without sample size re-estimation.

email: jingzhou@live.unc.edu

EVALUATING TYPE I ERROR RATE IN DESIGNING BAYESIAN ADAPTIVE CLINICAL TRIALS: A CASE STUDY
Manuela Buzoianu*, U.S. Food and Drug Administration

Bayesian methods offer increased flexibility for adaptive design, being able to combine prior information and accumulated data during a clinical trial in developing pre-specified decision rules regarding enrollment, futility, or success. The concern about controlling type I error rate in Bayesian adaptive clinical trials needs to be addressed, in particular in the regulatory setting. One Bayesian design method that adaptively determines the sample size is based on evaluating the predictive probabilities of eventual trial success at interim looks and employs the use of a probability model for the transitions from interim states to final stage outcome. The transition probabilities are often given informative priors. In large clinical trials these priors may have minimal impact on the overall type I error rate. A simulation study is conducted to assess how sensitive the operating characteristics, in particular the type I error rate, are to the transition probability prior distributions at various values of actual transition probabilities.

email: manuelabuzoianu@yahoo.com

A CONDITIONAL ERROR RATE APPROACH TO ADAPTIVE ENRICHMENT TRIAL DESIGNS
Brent R. Logan*, Medical College of Wisconsin

Adaptive enrichment clinical trial designs have been proposed to allow mid-trial flexibility in targeting a subpopulation identified as most likely to respond to treatment. They include testing of the treatment effect in both the overall population and the subpopulation, an interim assessment of whether to continue with the full population or restrict future eligibility to the subpopulation, and possible reassessment of the sample size. Multiple testing adjustment is needed to control the type I error rate due to the multiple populations being considered as candidates for enrollment in the second stage. Several strategies have been proposed for this adjustment, using principles of closed testing procedures applied to prespecified weighted combinations of the stages of data. Here we propose a conditional error rate approach applied to each intersection null hypothesis in the closed test procedure, based on a planned multiple testing strategy assuming the study population is not changed in the second stage. If the enrollment is restricted to the subpopulation the second stage p-value is compared to the conditional error rate to determine whether there is a significant effect in the subpopulation. We compare this approach to other methods using a simulation study.

email: blogan@mcw.edu

A PHASE I BAYESIAN ADAPTIVE DESIGN TO SIMULTANEOUSLY OPTIMIZE DOSE AND SCHEDULE ASSIGNMENTS BOTH AMONG AND WITHIN PATIENTS
Jin Zhang*, University of Michigan
Thomas Braun, University of Michigan

In traditional schedule or dose-schedule finding designs, patients are assumed to receive their assigned dose-schedule combination throughout the trial even though the combination may be found to have an undesirable toxicity profile, which contradicts actual clinical practice. Since no systematic approach exists to optimize intra-patient dose-schedule assignment, we propose a Phase I clinical trial design that extends existing approaches that optimize dose and schedule solely among patients by incorporating adaptive variations to dose-schedule assignments within patients as the study proceeds. Our design is based on a Bayesian non-mixture cure rate model that incorporates multiple administrations each patient receives with the per-administration dose included as a covariate. Simulations demonstrate that our design identifies safe dose and schedule combinations as well as the traditional method that does not allow for intra-patient dose-schedule reassignments, but with a larger number of patients assigned to safe combinations.

email: zhjin@umich.edu

A SEMI-PARAMETRIC APPROACH FOR DESIGNING SEAMLESS PHASE II/III STUDIES WITH TIME-TO-EVENT ENDPOINTS
Fei Jiang*, Rice University
Yanyuan Ma, Texas A&M University
J. Jack Lee, University of Texas MD Anderson Cancer Center

We develop a seamless phase II/ III clinical trial design to evaluate the efficacy of new treatments. The endpoints for the phase II and phase III studies are the progression free survival (PFS) and the overall survival (OS), respectively. The main goal for the phase II part of the study is to drop the inefficacious treatments while sending the efficacious ones for further evaluation in the phase III part. The phase III part of the study focuses on evaluating the effectiveness (OS) of the selected treatments and making the final decision on whether the new treatments is better than the standard treatment or not. Since the long term response (OS) is largely unobservable in the phase II study, we incorporate the information from the short term endpoint (FPS) to make the inference. We propose a semi-parametric method to model the short term and long term survival times. Furthermore, we propose a score function imputation method for the estimation under censoring. The theoretic derivations and the numerical results show that the proposed estimators are consistent and more efficient than the least square estimators. We use the estimators in the clinical trial designs. We allow the early stopping of the trial due to the futility of the new treatments. The simulation studies show that the proposed designs can control both the type I and II errors as well as reduced the average sample sizes of the trial compared to the traditional design of a phase II study followed by a phase III study. The expected study duration for the seamless design is also shorter compared to the traditional design.

email: homebovine@hotmail.com

81. METHODS FOR SURVIVAL ANALYSIS

DOUBLY-ROBUST ESTIMATORS OF TREATMENT-SPECIFIC SURVIVAL DISTRIBUTIONS IN OBSERVATIONAL STUDIES WITH STRATIFIED SAMPLING
Xiaofei Bai*, North Carolina State University
Anastasios (Butch) Tsiatis, North Carolina State University
Sean M. O'Brien, Duke University

Observational studies are frequently conducted to compare the effects of two treatments on survival. For such studies we must be concerned about confounding; that is, there are covariates that affect both the treatment assignment and the survival distribution. With confounding the usual treatment-specific Kaplan-Meier estimator might be a biased estimator of the underlying treatment-specific survival distribution. This paper has two aims. In the first aim we use semiparametric theory to derive a doubly robust estimator of the treatment-specific survival distribution in cases where it is believed that all the potential confounders are captured. In cases where not all potential confounders have been captured one may conduct a substudy using a stratified sampling scheme to capture additional covariates that may account for confounding. The second aim is to derive a doubly-robust estimator for the treatment-specific survival distributions and its variance estimator with such a stratified sampling scheme. Simulation studies are conducted to show consistency and double robustness. These estimators are then applied to the data from the ASCERT study that motivated this research.

email: xbai3@ncsu.edu

IDENTIFYING IMPORTANT EFFECT MODIFICATION IN PROPORTIONAL HAZARD MODEL USING ADAPTIVE LASSO
Jincheng Shen*, University of Michigan
Lu Wang, University of Michigan

In many biomedical studies, effect modification is often of scientific interest to investigators. For example, it is important to evaluate the effect of gene-gene and gene-environment interactions on many complex diseases. When it involves a large number of genetic and environmental risk factors, identifying important interactions becomes challenging. We propose a modified adaptive LASSO method for Cox's proportional hazard

model to simultaneously fit a Cox regression model with right censored time to event outcome and select important interaction terms. The proposed method uses the penalized log partial likelihood with adaptively weighted L1 penalty on regression coefficients, and it enforces the heredity constraint automatically, that is, an interaction term can be selected if and only if the corresponding main terms are also included in the model. Asymptotic properties including consistency and rate of convergence are studied, and the proposed selection procedure is also shown to have the oracle properties with proper choice of regularization parameters. The two-dimensional tuning parameter is determined by generalized cross-validation. Numerical results on both simulation data and real data demonstrate that our method performs effectively and competitively.

email: jcshen@umich.edu

PROXIMITY OF WEIGHTED AND LINDLEY MODELS WITH ESTIMATION FROM CENSORED SAMPLES

Broderick O. Oluyede*, Georgia Southern University
Mutiso Fidelis, Georgia Southern University

Proximity of the Lindley distribution to the class of increasing failure rate (IFR) and decreasing failure rate (DFR) weighted distributions including transformed distributions with monotone weight functions are obtained. Maximum likelihood estimates of the parameters of the generalized Lindley distribution are presented from samples with type I right, type II doubly and type II progressively censored data. An empirical analysis using published censored data sets are given in which the generalized Lindley distribution fits the data better than other well known and commonly used parametric distributions in survival and reliability analysis.

email: boluyede@georgiasouthern.edu

PARAMETER ESTIMATION IN COX PROPORTIONAL HAZARD MODELS WITH MISSING CENSORING INDICATORS

Naomi C. Brownstein*, University of North Carolina, Chapel Hill
Eric Bair, University of North Carolina, Chapel Hill
Jianwen Cai, University of North Carolina, Chapel Hill
Gary Slade, University of North Carolina, Chapel Hill

In a prospective cohort study, examining all participants for incidence of the condition of interest may be prohibitively expensive. For example, the 'gold standard' for diagnosing temporomandibular disorder (TMD) is a clinical examination by an expert dentist. Examining all subjects in this manner is infeasible for large studies. Instead, it is common to use a cheaper examination to screen for possible incident cases and perform the 'gold standard' examination only on participants who screen positive. Unfortunately, subjects may leave the study before receiving the 'gold standard' examination. Within the framework of survival analysis, this results in missing censoring indicators. Motivated by the Orofacial Pain: Prospective Evaluation and Risk Assessment (OPPERA) study, we propose a method for parameter estimation in survival models with missing censoring indicators. We estimate the probability of being a case for those with no 'gold standard' examination using logistic regression. These predicted probabilities are used to generate multiple imputations of each missing case status and estimate the hazard ratios associated with each putative risk factor. The variance introduced by the procedure is estimated using multiple imputation. We simulate data with missing censoring indicators and show that our method has similar or better performance than the competing methods.

email: nbrownst@email.unc.edu

NONPARAMETRIC BAYES ESTIMATION OF GAP-TIME DISTRIBUTION WITH RECURRENT EVENT DATA

AKM F. Rahman*, University of South Carolina, Columbia
James Lynch, University of South Carolina, Columbia
Edsel A. Pena, University of South Carolina, Columbia

Nonparametric Bayes estimation of the gap-time survivor function governing the time to occurrence of a recurrent event in the presence of censoring is considered. Denote the successive gap-times of a recurrent event by { T_ij , k=1, 2, 3, & } and the end of monitoring time by Ä_i . Assume that T_ij ~ F, Ä_i~G , and T_ij and Ä_i are independent. In our Bayesian approach, F has a Dirichlet process prior with parameter ±, a non-null finite measure on (R_+,Ã(R_+)). We derive nonparametric Bayes (NPB) and empirical Bayes (NPEB) estimators of the survivor function of F =1-F. The resulting NPB estimator of F extends the Bayes estimator of F =1-F in Susarla and Van Ryzin (1976) based on a single-event right-censored data. The PL-type nonparametric estimator of F based on recurrent event data presented in Pena et al. (2001) is also a limiting case of the NPB estimator, obtained by letting ±'0 . Through simulation studies, we demonstrate that the PL-type estimator has smaller bias but higher root-mean-squared errors (RMSE) than those of the NPB and the NPEB estimators. Even in the case of a misspecified prior measure parameter ±, the NPB and NPEB estimators have smaller RMSE than the PL-type estimator, indicating robustness of the NPB and NPEB estimators. In addition, NPB and NPEB estimator is smoother than the PL-type estimator.

email: rahmana@email.sc.edu

BAYESIAN REGRESSION ANALYSIS OF MULTIVARIATE INTERVAL-CENSORED DATA

Xiaoyan Lin*, University of South Carolina
Lianming Wang, University of South Carolina

We propose a frailty semiparametric Probit model for multivariate interval-censored data. We allow different correlations for different pairs of responses by using a shared normal frailty combined with multiple different coefficients among the failure types. The correlations in terms of Kendall's tau have simple and explicit forms. The regression parameters have a nice interpretation as the marginal and conditional covariate effects have a determined relationship via the variances of the frailties. Monotone splines are used to model the nonparametric functions in the frailty Probit model. An easy-to-implement MCMC algorithm is developed for the posterior computation. The proposed method is evaluated using simulation and is illustrated by a real-life data about multiple sexually transmitted infections.

email: lin9@mailbox.sc.edu

A BAYESIAN SEMIPARAMETRIC APPROACH FOR THE EXTENDED HAZARDS MODEL

Li Li*, University of South Carolina
Timothy Hanson, University of South Carolina

In this paper we consider the extended hazards model of Chen and Jewell (2001). Despite including three popular survival models as special cases -- proportional hazards, accelerated failure time, and accelerated hazards -- this model has received very little attention in the literature due to a highly challenging likelihood, espeically when the baseline hazard is left to be completely arbitrary. We consider a Bayesian semiparametric approach based on B-splines that is centered at a given parametric hazard family through a parametric quantile function. The resulting model is very smooth, yet highly flexible, and through strategic blocking of model parameters and clever data augmentation, the MCMC is reasonably straightforward. The basic model is developed and illustrated, including tests for proportional hazards, accelerated failure time, and accelerated hazards carried out through approximate Bayes factors.

email: lil@email.sc.edu

82. META-ANALYSIS

ADAPTIVE FUSSED LASSO IN META LONGITUDINAL STUDIES

Fei Wang*, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health and Wayne State University
Lu Wang, University of Michigan
Peter X.-K. Song, University of Michigan

We develope a screening procedure to detect parameter homogeneity and reduce model complexity in the process of data merging. We consider the longitudinal marginal model for merged studies, in which the classical hypothesis testing approach to evaluating all possible subsets of common regression parameters can be combinatorially complex and computationally prohibitive. We develop a regularization method that can overcome this difficulty by applying the idea of adaptive fused lasso in that restrictions are imposed on differences of pairs of parameters between studies. The selection procedure will automatically detect common parameters across all or subsets of studies. Through simulation studies we show that the proposed method performs well to consistently identify common parameters.

email: wafei@umich.edu

EFFICIENT META-ANALYSIS OF HETEROGENEOUS STUDIES USING SUMMARY STATISTICS

Dungang Liu*, Yale University
Regina Liu, Rutgers University
Minge Xie, Rutgers University

Meta-analysis has been widely used to synthesize evidence from multiple studies. A practical and important issue in meta-analysis is that the studies are often found heterogeneous, due to different study populations, designs and outcomes. In the presence of heterogeneous studies, we may encounter the problem that the parameter of interest is not estimable in some of the studies. Consequently, such studies are typically excluded from the conventional meta-analysis, which can lead to non-negligible loss of information. In this paper, we propose an efficient meta-analysis approach that can incorporate all the studies in the analysis. This approach combines confidence density functions obtained from each of the studies. It can integrate direct as well as indirect evidence in the analysis. Under a general likelihood inference framework, we show that our approach is asymptotically as efficient as the maximum likelihood approach using individual participant data (IPD) from all the studies. But unlike the IPD analysis, our approach only uses summary statistics and does not require individual-level data. In fact, our approach provides a unifying treatment for combining summary statistics, and it subsumes several existing meta-analysis methods. In addition, our approach is robust against misspecification of the working covariance structure of the parameter estimates. These desirable properties are confirmed numerically in the analysis of the simulated data from a randomized clinical trials setting and the real data on flight landing retrieved from the Federal Aviation Administration (FAA).

email: dungang.liu@yale.edu

GENERAL FRAMEWORK FOR META-ANALYSIS FOR RARE VARIANTS ASSOCIATION STUDIES

Seunggeun Lee*, Harvard School of Public Health
Tanya Teslovich, University of Michigan
Michael Boehnke, University of Michigan
Xihong Lin, Harvard School of Public Health

We propose a general statistical framework for meta-analysis for group-wise rare variant association tests. In genomic studies, meta-analysis has been widely used in single marker analysis to increase statistical power by combining data from different studies. Currently, an active area of research is joint analysis of rare variants, to discover of sets of trait-associated variants even when studies are underpowered to detect rare variants individually. To facilitate meta-analysis of this new class of association tests, we have developed a novel meta-analysis method that extends popular gene/region based rare variants tests, such as burden test, sequence kernel association test (SKAT) and more recent optimal unified test (SKAT-O), to the meta-analysis framework. The proposed method uses gene-level summary statistics to conduct association tests and is flexible enough to incorporate different levels of heterogeneity of genetic effect across cohorts, and even group-wise heterogeneity for multi-ethnic studies. Furthermore, the proposed method is as powerful as joint analysis of all individual level genetic data. Extensive simulations with varying levels of heterogeneity and real data analysis demonstrate the superior performance of our method.

email: sglee@hsph.harvard.edu

NONPARAMETRIC INFERENCE FOR META ANALYSIS WITH A SET OF FIXED, UNKNOWN STUDY PARAMETERS

Brian Claggett*, Harvard School of Public Health
Min-ge Xie, Rutgers University
Lu Tian, Stanford University
Lee-Jen Wei, Harvard School of Public Health

Meta-analysis is a valuable tool for combining information from independent studies. However, most common meta-analysis techniques rely on distributional assumptions that are difficult, if not impossible, to verify. For instance, in the commonly used fixed-effects and random-effects models, we take for granted that the underlying study parameters are either exactly the same across individual studies or that they are random samples from a population under a parametric distributional assumption. In this paper, we present a new framework for summarizing information obtained from multiple studies and make inference that is not dependent on any distributional assumption for the study-level unknown, fixed parameters. Specifically, we draw inferences about, for example, the quantiles of this set of parameters using study-specific summary statistics. This type of problem is quite challenging (Hall and Miller, 2010). We utilize a novel resampling method via confidence distributions of study-specific parameters to construct confidence intervals for the above quantiles. We justify the validity of the interval estimation procesure asymptotically and compare the new procesure with the standard bootstrapping method. We also illustrate our proposal with the data from a recent meta analysis of the treatment effect from an antioxidant on the prevention of contrast-induced nephropathy.

email: bclagget@hsph.harvard.edu

TOWARDS PATIENT-CENTERED NETWORK META-ANALYSIS OF RANDOMIZED CLINICAL TRIALS WITH BINARY OUTCOMES: REPORTING THE PROPER SUMMARIES

Jing Zhang*, University of Minnesota
Bradley P Carlin, University of Minnesota
James D. Neaton, University of Minnesota
Guoxing G. Soon, U.S. Food and Drug Administration
Lei Nie, U.S. Food and Drug Administration
Robert Kane, University of Minnesota
Beth A. Virnig, University of Minnesota
Haitao Chu, University of Minnesota

In the absence of sufficient data directly comparing multiple treatments, indirect comparisons using network meta-analyses (NMA) across trials can potentially provide useful information to guide the use of treatments. Under current contrast-based methods of binary outcomes, the proportion of responders for each treatment and risk differences are not provided. Most NMAs only report odds ratios which may be misleading when events are common. A novel Bayesian hierarchical model, developed from a missing data perspective, is used to illustrate how treatment-specific event proportions, risk differences (RD) and relative risks (RR) can be computed in NMAs. We first compare our approach to alternative methods using two hypothetical NMAs, and then use a published NMA on new-generation antidepressants to illustrate the improved reporting of NMAs. In the hypothetical NMAs, our approach outperforms current methods in terms of bias. In the published NMA, the outcomes were common with proportions ranging from 0.21 to 0.62. In addition, differences in the magnitude of relative treatment effects and statistical significance of several pairwise comparisons from previous report could lead to different treatment recommendations. In summary, the proposed NMA method can accurately estimate treatment-specific event proportions, RDs, and RRs, and is recommended in practice.

email: jingzhang2773691@gmail.com

STATISTICAL CHARACTERIZATION AND EVALUATION OF MICROARRAY META-ANALYSIS METHODS: A PRACTICAL APPLICATION GUIDELINE

Lun-Ching Chang*, University of Pittsburgh
Hui-Min Lin, University of Pittsburgh
George C. Tseng, University of Pittsburgh

As high-throughput genomic technologies become more accurate and affordable, an increasing number of data sets have accumulated in the public domain and genomic information integration and meta-analysis have become routine in biomedical research. In this paper, we focus on microarray meta-analysis, where multiple microarray studies with relevant biological hypotheses are combined in order to improve candidate marker detection. Many methods have been developed and applied, but their performance and properties have only been minimally investigated. There is currently no clear conclusion or guideline as to the proper choice of a meta-analysis method given an application. Here we perform a comprehensive comparative analysis for twelve microarray meta-analysis methods through simulations and six large-scale applications using four evaluation criteria. We elucidate hypothesis settings behind the methods and further apply multi-dimensional scaling (MDS) and an entropy measure to characterize the meta-analysis methods and data structure, respectively. The aggregated results provide an insightful and practical guideline to the choice of the most suitable method in a given application.

email: lunching@gmail.com

UNCONFOUNDING THE CONFOUNDED: ADJUSTING FOR BATCH EFFECTS IN COMPLETELY CONFOUNDED DESIGNS IN GENOMIC STUDIES

W. Evan Johnson*, Boston University School of Medicine
Timothy M. Bahr, University of Iowa School of Medicine

Batch effects are often observed across multiple batches of high-throughput data. Supervised and unsupervised methods have previously been developed to account for simple batch effects forexperiments where each batch contains multiple control and treatment samples. However, no method has been designed for data sets where the control samples are completely contained in one set of batches and the treatment samples are contained in another set of batches. Data sets that are completely confounded in this manner are generally discarded due to lack of identifiability between treatment and batch effects. Here we propose a method that uses a rank test and an Empirical Bayes framework to model systematically varying batch effects in completely confounded designs or in batches that contain a single sample. We then are able to adjust data sets for batch effects and accurately estimate treatment effects. We illustrate the robustness of our method through a simulation study and an application to a real multiple-batch data set and show that our method are useful, justifiable, and very robust in practice. The method is implemented in the 'ComBat' function in the 'sva' Bioconductor package.

email: wej@bu.edu

83. STATISTICAL METHODS IN CANCER APPLICATIONS

RECLASSIFICATION OF PREDICTIONS FOR COMPARING RISK PREDICTION MODELS

Swati Biswas*, University of Texas, Dallas
Banu Arun, University of Texas MD Anderson Cancer Center
Giovanni Parmigiani, Dana Farber Cancer Institute and Harvard School of Public Health

Risk prediction models play an important role in prevention and treatment of several diseases. Models that are in clinical use are often refined and improved. In many instances, the most efficient way to improve a 'successful' model is to identify subgroups of populations for which a specific biological rationale exists and tailor the improved model to those subjects, an approach especially in line with personalized medicine. At present, we lack statistical tools to evaluate improvements targeted to specific sub-groups. Here we propose simple tools to fill this gap. First, we extend a recently proposed measure, Integrated Discrimination Improvement, using a linear model with covariates representing the sub-groups. Next, we develop graphical and numerical tools that compare reclassification of two models but focusing only on those subjects for whom the two models reclassify differently. We apply these approaches to the genetic risk prediction model for breast cancer BRCAPRO, using clinical data from MD Anderson Cancer Center. We also conduct a simulation study to investigate properties of the new reclassification measure and compare it with currently used measures. Our results show that the proposed tools can successfully uncover sub-group specific model improvements.

email: swati.biswas@utdallas.edu

UPDATING EXISTING RISK PREDICTION TOOLS FOR NEW BIOMARKERS

Donna P. Ankerst*, Technical University, Munich
Andreas Boeck, Technical University, Munich

Online risk prediction tools for common cancers are now easily accessible and widely used by patients and doctors for informed decision-making concerning screening. A practical problem is as cancer research moves forward and new biomarkers are discovered, there is a need to update the risk algorithms to include them. Typically, the new markers cannot be retrospectively measured on the same study participants used to develop the original prediction tool, necessitating the merging of a separate study of different participants, which may be much smaller in sample size and of a different design. This talk reports on the application of Bayes rule for updating risk prediction tools to include a set of biomarkers measured in an external study to the original study used to develop the risk prediction tool. The procedure is illustrated in the context of updating the online Prostate Cancer Prevention Trial Risk Calculator (PCPTRC) to incorporate the new markers %freePSA and [-2]proPSA and single nucleotide polymorphisms recently identified through genomewide association studies. The updated PCPTRC models are compared to the existing PCPTRC through validation on an external data set, based on discrimination, calibration and clinical net benefit metrics.

email: ankerst@ma.tum.de

PARAMETRIC AND NON PARAMETRIC ANALYSIS OF COLON CANCER

Venkateswara Rao Mudunuru*, University of South Florida
Chris P. Tsokos, University of South Florida

The object of the present study is to perform statistical analysis of malignant colon tumor with the tumor size being the response variable. We determined that the tumor sizes of whites, African Americans and other races are statistically different. The probability distribution that characterize the behavior of the response variable was obtained along with the confidence limits. The malignant tumor size as a function of age was partitioned into three significant age intervals and the mathematical function that characterizes the size of the tumor as a function of age was determined for each age interval.

email: vmudunur@mail.usf.edu

ASSESSING INTERACTIONS FOR FIXED-DOSE DRUG COMBINATIONS IN TUMOR XENOGRAFT STUDIES

Jianrong Wu*, St. Jude Children's Research Hospital
Lorriaine Tracey, St. Jude Children's Research Hospital
Andrew Davidoff, National University of Singapore

Statistical methods for assessing the joint action of compounds administered in combination have been established for many years. However, there is little literature available on assessing the joint action of fixed-dose drug combinations in tumor xenograft experiments. Here an interaction index for fixed-dose two-drug combinations is proposed. Furthermore, a regression analysis is also discussed. Actual tumor xenograft data were analyzed to illustrate the proposed methods.

email: jianrong.wu@stjude.org

KERNELIZED PARTIAL LEAST SQUARES FOR FEATURE REDUCTION AND CLASSIFICATION OF GENE MICROARRAY DATA

Walker H. Land, Binghamton University
Xingye Qiao*, Binghamton University
Daniel E. Margolis, Binghamton University
William S. Ford, Binghamton University
Christopher T. Paquette, Binghamton University
Joseph F. Perez-Rogers, Binghamton University
Jeffrey A. Borgia, Rush University Medical Center
Jack Y. Yang, Harvard Medical School
Youping Deng, Rush University Medical Center

The primary objectives of this paper are: 1.) to apply Statistical Learning Theory, specifically Partial Least Squares (PLS) and Kernelized PLS, to the universal 'feature-rich/case-poor' (also known as 'large p small n', or 'high-dimension, low-sample size') microarray problem by eliminating those features (or probes) that do not contribute to the 'best' chromosome bio-markers for lung cancer, and 2.) quantitatively measure and verify (by an independent means) the efficacy of this PLS process. A secondary objective is to integrate these significant improvements in diagnostic and prognostic biomedical applications into the clinical research arena. Statistical learning techniques such as PLS and Kernelized PLS can effectively address difficult problems with analyzing biomedical data such as microarrays. The combinations with established biostatistical techniques demonstrated in this paper allow these methods to move from academic research and into clinical practice.

email: qiao@math.binghamton.edu

GENE EXPRESSION DECONVOLUTION IN HETEROGENOUS TUMOR SAMPLES

Jaeil Ahn, University of Texas MD Anderson Cancer Center
Giovanni Parmigiani, Dana Farber Cancer Institute and Harvard School of Public Health
Ying Yuan, University of Texas MD Anderson Cancer Center
Wenyi Wang* University of Texas MD Anderson Cancer Center

Clinically derived tumor tissues are often times made of both cancer and normal stromal cells. The expression measures of these samples are therefore partially derived from the non-tumor cells. This may explain why some previous studies have identified only a fraction of differentially expressed genes between tumor and normal samples. What makes the in silico estimation of mixture components more difficult is that the percentage of normal cells varies from one tissue sample to another. Until recently, there has been limited work on statistical methods development that accounts for tumor heterogeneity in gene expression data. To this end, we have developed a likelihood-based method to estimate, in each tumor sample, the normal cell fractions (i.e. level of stromal contamination), as well as cancer cell-specific gene expressions. We illustrate the performance of our model in synthetic as well as in real clinical data.

email: wwang7@mdanderson.org

eQTL MAPPING USING RNA-seq DATA FROM CANCER PATIENTS

Wei Sun*, University of North Carolina, Chapel Hill

We study gene expression QTL (eQTL) mapping using RNA-seq data from The Cancer Genome Atlas Project. In particular, we will study the association between germline DNA mutation or somatic DNA mutations with gene expression in tumor tissue. New statistical methods will be developed to assess both allele-specific (cis-acting)-eQTL and RNA-isoform-specific eQTL.

email: weisun@email.unc.edu

84. RECENT METHODOLOGICAL ADVANCES IN THE ANALYSIS OF CORRELATED DATA

AN IMPROVED QUADRATIC INFERENCE APPROACH FOR THE MARGINAL ANALYSIS OF CORRELATED DATA

Philip M. Westgate*, University of Kentucky

Generalized estimating equations (GEE) are commonly employed for the analysis of correlated data. However, the Quadratic Inference Function (QIF) method is increasing in popularity due to its multiple theoretical advantages over GEE. Our focus is that the QIF method is more efficient than GEE when the working covariance structure for the data is misspecified. However, for finite-sample sizes, the QIF method may not perform as well as GEE due to the use of an empirical covariance matrix in its estimating equations. We discuss adjustments to the estimation procedure and the covariance matrix that improve the QIF's performance, relative to GEE, in finite-samples. We then discuss theoretical and realistic advantages and disadvantages of the use of additional or alternative estimating equations within the QIF approach, and propose a method to choose the estimating equations that will result in the least variable parameter estimates, and thus improved inference.

email: philip.westgate@uky.edu

ASSESSING VARIANCE COMPONENTS IN MULTILEVEL LINEAR MODELS USING APPROXIMATE BAYES FACTORS

Ben Saville*, Vanderbilt University

Deciding which predictor effects may vary across subjects is a difficult issue. Standard model selection criteria and test procedures are often inappropriate for comparing models with different numbers of random effects due to constraints on the parameter space of the variance components. Testing on the boundary of the parameter space changes the asymptotic distribution of some classical test statistics and causes problems in approximating Bayes factors. We propose a simple approach for assessing variance components in multilevel linear models using Bayes factors. We scale each random effect to the residual variance and introduce a parameter that controls the relative contribution of each random effect free of the scale of the data. We integrate out the random effects and the variance components using closed form solutions. The resulting integrals needed to calculate the Bayes factor are low-dimensional integrals lacking variance components and can be efficiently approximated with Laplace's method. We propose a default prior distribution on the parameter controlling the contribution of each random effect and conduct simulations to show that our method has good properties for model selection problems. Finally, we illustrate our methods on data from a clinical trial of patients with bipolar disorder and on a study of racial/ethnic disparities of infant birthweights.

email: b.saville@vanderbilt.edu

MERGING LONGITUDINAL OR CLUSTERED STUDIES: VALIDATION TEST AND JOINT ESTIMATION

Fei Wang, Wayne State University
Lu Wang*, University of Michigan
Peter X.K. Song, University of Michigan

Merging data from multiple studies has been widely adopted in biomedical research. In this paper, we consider two major issues related to merging longitudinal datasets. We first develop a rigorous hypothesis testing procedure to assess the validity of data merging, and then propose a flexible joint estimation procedure that enables us to analyze merged data and to account for different within-subject correlations and follow-up schedules in different studies. We establish large sample properties for the proposed procedures. We compare our method with meta analysis and generalized estimating equation and show that our test provides robust control of type I error against both misspecification of working correlation structures and heterogeneous dispersion parameters. Our joint estimating procedure leads to an improvement in estimation efficiency on all regression coefficients after data merging is validated.

email: luwang@umich.edu

MODELING THE DISTRIBUTION OF PERIODONTAL DISEASE WITH A GENERALIZED VON MISES DISTRIBUTION

Thomas M. Braun*, University of Michigan
Samopriyo Maitra, University of Michigan

Periodontal disease is a common cause of tooth loss in adults. The severity of periodontal disease is usually quantified based upon the magnitudes of several tooth-level clinical parameters, the most common of which is clinical attachment level (CAL). Recent clinical studies have presented data on the distribution of periodontal disease in hopes of providing information for localized treatments that can reduce the prevalence of periodontal disease. However, these findings have been descriptive without consideration of statistical modeling for estimation and inference. To this end, we visualize the mouth as a circle and the teeth as points located on the circumference of the circle to allow the use of circular statistical methods to determine the mean locations of diseased teeth. We assume the directions of diseased teeth, as determined by their tooth averaged CAL values, to be observations from a Generalized von Mises distribution. Because multiple teeth from a subject are correlated, we use a bias-corrected generalized estimating equation approach to obtain robust variance estimates for our parameter estimates. Via simulations of data motivated from an actual study of periodontal disease, we demonstrate that our methods have excellent performance in the moderately small sample sizes common to most periodontal studies.

email: tombraun@umich.edu

85. FRONTIERS IN STATISTICAL GENETICS AND GENOMICS

BAYESIAN INFERENCE OF SPATIAL ORGANIZATIONS OF CHROMOSOMES

Ming Hu, Harvard University
Ke Deng, Harvard University
Zhaohui Qin, Emory University
Bing Ren, University of California, San Diego
Jun S. Liu*, Harvard University

Knowledge of spatial chromosomal organizations is critical for the study of transcriptional regulation and other nuclear processes in the cell. Recently, chromosome conformation capture (3C) based technologies, such as Hi-C and TCC, have been developed to provide a genome-wide, three-dimensional (3D) view of chromatin organization. Here we describe a novel Bayesian probabilistic approach, BACH, to infer the consensus 3D chromosomal structure. In addition, we describe a variant algorithm BACH-MIX to study the structural variations of chromatin in a cell population. Applying BACH and BACH-MIX to a high resolution found that most local genomic regions exhibit homogeneous Hi-C dataset generated from mouse embryonic stem cells, we 3D chromosomal structures. We further constructed a model for the spatial arrangement of chromatin, which reveals structural properties associated with euchromatic and heterochromatic regions in the genome. We observed strong associations between structural properties and several genomic and epigenetic features of the chromosome. Using BACH-MIX, we further found that the structural variations of chromatin are correlated with these genomic and epigenetic features. Our results demonstrate that BACH and BACH-MIX have the potential to provide new insights into the chromosomal architecture of mammalian cells.

email: jliu1600@gmail.com

MICROBIOME, METAGENOMICS AND HIGH DIMENSIONAL COMPOSITION DATA

Hongzhe Li*, University of Pennsylvania

With the development of next generation sequencing technology, researchers have now been able to study the microbiome composition using direct sequencing, whose output are bacterial taxa counts for each microbiome sample. One goal of microbiome studies is to associate the microbiome composition with environmental covariates or clinical outcomes, including (1) identification of the biological/environemental factors that are associated with bacterial compositions; (2) identification of the bacterial taxa that are associated with clinical outcomes. Statistical models to address these problems need to account for the high-dimensional , sparse and compositional nature of the data. In addition, the prior phylogenetic tree among the bacterial species provides useful information on bacterial phylogeny. In this talk, I will present several statistical methods we developed for analyzing the bacterial compositional data, including kernel-based regression, sparse Dirichlet-multinomial regression, compositional data regression and construction of bacterial taxa network based on compositional data. I demonstrate the methods using a data set that links human gut microbiome to diet intake in order to identify the micro-nutrients that are associated with the human gut microbiome and the bacteria that are associated with body mass index.

email: hongzhe@upenn.edu

DESIGNS AND ANALYSIS OF SEQUENCING STUDIES WITH TRAIT-DEPENDENT SAMPLING

Danyu Lin*, University of North Carolina, Chapel Hill

It is not economically feasible to sequence all study subjects in a large cohort. A cost-effective strategy is to sequence only the subjects with the extreme values of a quantitative trait. In the NHLBI Exome Sequencing Project, subjects with the highest or lowest values of BMI, LDL or blood pressures were selected for whole-exome sequencing. In the NHLBI Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) resequencing project, subjects with the highest values on one of twenty quantitative traits were selected for target sequencing, along with a random sample. Failure to account for such trait-dependent sampling can lead to severe inflation of type I error and substantial loss of power in the quantitative trait analysis. We present valid and efficient likelihood-based inference procedures under general trait-dependent sampling. Our methods can be used to perform quantitative trait analysis not only for the trait that is used to select subjects for sequencing but also for any other traits that are measured. We also investigate the relative efficiency of various sampling strategies. The proposed methods are demonstrated through simulation studies and the aforementioned NHLBI sequencing projects.

email: lin@bios.unc.edu

AN EMPIRICAL BAYESIAN FRAMEWORK FOR ASSESSMENT OF INDIVIDUAL-SPECIFIC RISK OF RECURRENCE

Kevin Eng, University of Wisconsin, Madison
Shuyun Ye, University of Wisconsin, Madison
Ning Leng, University of Wisconsin, Madison
Christina Kendziorski*, University of Wisconsin, Madison

Accurate assessment of an individual's risk for disease recurrence following an initial treatment has implications for patient health as it enables an individual to receive interventionary measures, potentially enabling a later recurrence, or preventing the recurrence altogether. Toward this end, we have developed an empirical Bayesian framework that combines measurements from high-throughput genomic technologies with medical records to estimate an individual's risk of a time-to-event phenotype. The framework has high sensitivity and specificity for predicting those at risk for ovarian cancer recurrence, is validated in independent data sets and experimentally, and is used to suggest alternative treatments.

email: kendzior@biostat.wisc.edu

86. BIG DATA: WEARABLE COMPUTING, CROWDSOURCING, SPACE TELESCOPES, AND BRAIN IMAGING

STATISTICAL CHALLENGES IN LARGE ASTRONOMICAL DATA SETS

Alexander S. Szalay*, Johns Hopkins University

Starting with the Sloan Digital Sky Survey, astronomy has started to collect very large data sets, covering a large part of the sky. Once these have been made publicly available, their analyses have created several nontrivial statistical and computational challenges. The talk will

discuss various problems related to spatial statistics, to the analysis a hundreds of thousands of galaxy spectra and novel ways of measuring galaxy distances. In most of these problems scalability is a primary concern. Also, with the cardinality of the data sets the dominant source of uncertainties are shifted from statistical errors to systematic ones. Robust subspace projection can be used to minimize some of the underlying systematics.

email: szalay@jhu.edu

VISUAL DATA MINING TECHNIQUES AND SOFTWARE FOR FUNCTIONAL ACTIGRAPHY DATA

Juergen Symanzik*, Utah State University
Abbass Sharif, Utah State University

Actigraphy, a technology for measuring a subject's overall activity level almost continuously over time, has gained a lot of momentum over the last few years. An actigraph, a watch-like device that can be attached to the wrist or ankle of a subject, uses an accelerometer to measure human movement every minute or even every 15 seconds. Actigraphy data are often treated as functional data. In this talk, we will present recently developed multivariate visualization techniques for actigraphy data. These techniques can be accessed via a user-friendly web interface that builds on an R package using an object-oriented model design that allows for fast modifications and extensions.

email: symanzik@math.usu.edu

AUTOMATIC SEGMENTATION OF LESIONS IN A LARGE LONGITUDINAL COHORT OF MULTIPLE SCLEROSIS SUBJECTS

Ciprian Crainiceanu*, Johns Hopkins University

Multiple sclerosis (MS) is a chronic immune-mediated disease of the central nervous system that results in significant disability and mortality. MS is often clinically diagnosed and characterized using visual inspection of brain images that contain lesions in the magnetic resonance imaging (MRI) modalities of brain tissue. These lesions appear at different times and locations and have different shapes and sizes. Visually identifying and delineating these lesions is time consuming, costly, and prone to inter- and intra-observer variability. We will present two methods for automatically identifying lesions and tracking them over time. The first method, Oasis, is focused on estimating the location of all voxels that may be part of a lesion. The second one, SubLIME, is dedicated to identifying those voxels that have become part of a new or enlarging lesion between two visits. The combination of Oasis and SubLIME is now the standard pipeline for automatic MS lesion segmentation. Methods for analyzing the lesion movies are also introduced. The presentation will focus on the important statistical problems and the solutions associated with a very complex spatio-temporal scientific problem. The data set is Terabyte-sized and comes from a multi-year longitudinal study of multiple sclerosis subjects conducted at NIH.

email: ccrainic@jhsph.edu

eBIRD: STATISTICAL MODELS FOR CROWDSOURCED BIRD DATA

Daniel Fink*, Cornell University

Effective management of bird species across their ranges requires knowledge of where the species are living: their distributions and habitat associations. Often, detailed data documenting a species' distribution is not available for the entire region of interest, particularly for widely distributed species. To meet this challenge ecologists harness the efforts of large numbers of volunteers to collect broad-scale species monitoring data. In this presentation we describe the analysis of the crowdsourced bird observation data collected by eBird (http://www.ebird.org), to study continent-wide inter-annual migrations of North American birds. This data set contains over 3 million species checklists at over 450,000 unique locations within the continental U.S. The modeling challenge is to facilitate spatiotemporal pattern discovery with sparse, noisy data across a wide variety of species' migration dynamics. To do this we developed a simple mixture model for non-stationary spatiotemporal processes, SpatioTemporal Exploratory Models (STEMs). Using an allocation from XSEDE we calculated weekly species distribution estimates for over 200 bird species for the 2013 State of The Birds Report, a national conservation report. Ecologists are using these results to study how local-scale ecological processes vary across a species range, through time, and between species.

email: df36@cornell.edu

87. NOVEL DEVELOPMENTS IN THE CONSTRUCTION AND EVALUATION OF RISK PREDICTION MODELS

RISK ASSESSMENT WITH TWO PHASE STUDIES

Tianxi Cai*, Harvard University

Identification of novel biomarkers for risk assessment is important for both effective disease prevention and optimal treatment recommendation. Discovery relies on the precious yet limited resource of stored biological samples from large prospective cohort studies. Two-phase sampling designs provide a cost-effective tool in the context of biomarker evaluation. Existing methods focus on making efficient inference on relative hazard parameters from the Cox regression model. When the Cox model fails to hold, the resulting risk estimates may have unsatisfactory prediction accuracy. In this talk, we describe a novel approach to derive a robust risk score for prediction under a nnon-parametric transformation model. Taking an IPW approach, we propose a weighted objective function for estimating the model parameters, which directly relates to a type of C-statistic for survival outcomes. Hence regardless of model adequacy, the proposed procedure will yield a sensible composite risk score for prediction. A major obstacle for making inference under two-phase studies is due to the correlation induced by the finite population sampling. Standard procedures such as the bootstrap cannot be directly used for variance estimation. We propose a novel resampling procedure to derive confidence intervals for the model parameters.

email: tcai@hsph.harvard.edu

PROJECTING POPULATION RISK WITH COHORT DATA: APPLICATION TO WHI COLORECTAL CANCER DATA

Dandan Liu*, Vanderbilt University
Yingye Zheng, Fred Hutchinson Cancer Research Center
Li Hsu, Fred Hutchinson Cancer Research Center

Accurate and individualized risk prediction is valuable for successful management of chronic diseases such as cancer and cardiovascular diseases. Large cohort studies provide valuable resources for building risk prediction models, as the risk factors are collected at the baseline and subjects are followed over time until the occurrence of diseases or termination of the study. However, many cohorts are assembled for particular purposes, and hence their baseline risk may differ from the general population. Moreover for rare diseases the baseline risk may not be estimated reliably based on cohort data only. In this paper, we propose to make use of external disease incidence rate for estimating the baseline risk, which increases both efficiency and robustness. We proposed two sets of estimators and established the asymptotic distributions for both of them. Simulation results show that the proposed estimators are more efficient than the methods that do not utilize the external incidence rates. When the baseline hazard function of the cohort differs from the target population, the proposed estimators have less bias than the cohort-based Breslow estimators. We applied the method to a large cohort study, the Women's Health Initiative, for estimating colorectal cancer risk.

email: dandan.liu@vanderbilt.edu

EXTENSIONS OF CRITERIA FOR EVALUATING RISK PREDICTION MODELS FOR PUBLIC HEALTH APPLICATIONS

Ruth M. Pfeiffer*, National Cancer Institute, National Institutes of Health

We recently proposed two novel criteria to assess the usefulness of risk prediction models for public health applications. The proportion of cases followed, PCF(p), is the proportion of individuals who will develop disease who are included in the proportion p of individuals in the population at highest risk. The proportion needed to follow-up, PNF(q), is the proportion of the general population at highest risk that one needs to follow in order that a proportion q of those destined to become cases will be followed (Pfeiffer and Gail, 2011). Here, we introduce two new criteria by integrating PCF and PNF over a range of values of q or p to obtain iPCF, the integrated PCF, and iPNF, the integrated PNF. We estimate iPCF and iPNF based on observed risks in a population alone assuming that the risk model is well calibrated. We also propose and study estimates of PCF, PNF, iPCF and iPNF from case control data with known outcome prevalence and from cohort data, with baseline covariates and observed health outcomes. These estimates are consistent even the risk models are not well calibrated. We study the efficiency of the various estimates and propose tests for comparing two risk models, evaluated in the same validation data.

email: pfeiffer@mail.nih.gov

EVALUATING RISK MARKERS UNDER FLEXIBLE SAMPLING DESIGN

Yingye Zheng*, Fred Hutchinson Cancer Research Center
Tianxi Cai, Harvard School of Public Health
Margaret Pepe, Fred Hutchinson Cancer Research Center

Validation of a novel risk prediction model is often conducted using data from a prospective cohort study. Such design allows the calculation of distribution of risk for subjects with good and bad outcomes and related summary indices that characterize the predictive capacity of a risk model. We propose methods to calculate risk distributions and a wide variety of prediction indices when outcomes are censored failure times. The estimation procedures accommodate not only a full prospective cohort study design, but also two-phase study designs. In particular this talk will focus in depth a novel nested case-control design that is commonly undertaken in practice involves sampling until quotas of eligible cases and controls are identified. Different two-phase design options will be compared in terms of statistical efficiency and practical considerations in the context of risk model evaluation.

email: yzheng@fhcrc.org

88. SAMPLE SIZE PLANNING FOR CLINICAL DEVELOPMENT

THE USE OF ADAPTIVE DESIGNS IN THE EFFICIENT AND ACCURATE IDENTIFICATION OF EFFECTIVE THERAPIES

Scott S. Emerson*, University of Washington

Many clinical investigators have decried the high cost of drug development and the low rate of 'positive' studies among phase 3 clinical trials. There have been several initiatives to try to use adaptive designs to speed up this process. In its most general form, an adaptive design uses interim estimates of treatment effect to modify such trial parameters as maximal sample size, eligibility criteria, treatment parameters, or definition of outcomes. In this talk I discuss possible roles of adaptive designs in accelerating the 'drug discovery' process. In particular I focus on how the key distinction between screening and confirmatory trials plays in the phased investigation of new treatments, and the role that sequential and other adaptive clinical trial designs can play at each phase in increasing the number of effective therapies detected with limited resources.

email: semerson@uw.edu

SAMPLE SIZE RE-ESTIMATION BASED UPON PROMISING INTERIM RESULT: FROM 'LESS WELL UNDERSTOOD' TO 'WELL ACCEPTED'

Joshua Chen*, Merck

In its recent draft guidance document on adaptive designs for clinical trials, the FDA categorizes sample size re-estimation methods based upon unblinded interim result as 'less well understood', which has been interpreted by some clinical trialists as a stop signal for continued effort. There is much potential in these sample size adaptive methods to help improve the efficiency of clinical trials by reducing failure rate in costly late stage trials. Continued research to further understand the statistical characteristics, development of operational tools to implement the designs, and experiences from applications in the real world can be helpful to make these 'less well understood' methods well accepted by the clinical trial community. In this talk, I will discuss some sample size re-estimation methods with focus on the 50% conditional power approach, where the investigators have an added option to increase sample size without making any modification to the originally planned analyses if the interim result is promising (i.e., conditional power>50%) but there is an increased risk of failing. Experience with real clinical trials utilizing this sample size adaptation method will be shared.

email: joshua_chen@merck.com

SAMPLE SIZE EVALUATION IN CLINICAL TRIALS WITH CO-PRIMARY ENDPOINTS

Toshimitsu Hamasaki*, Osaka University Graduate School of Medicine
Takashi Sozu, Kyoto University School of Public Health
Tomoyuki Sugimoto, Hrosaki University Graduate School of Science and Technology
Scott Evans, Harvard School of Public Health

The determination of sample size and the evaluation of power are critical elements in the design of a clinical trial. If a sample size is too small then important effects may not be detected, while a sample size that is too large is wasteful of resources and unethically puts more participants at risk than necessary. Most commonly, a single primary endpoint is selected and is used as the basis for the trial design including sample size determination, interim monitoring, final analyses, and published reporting. However, many recent clinical trials have utilized co-primary endpoints potentially offering a more complete characterization of intervention effects. Use of co-primary endpoints creates complexities in the evaluation of power and sample size during trial design, specifically relating to control of Type II error when the endpoints are potentially correlated. We present an overview of an approach to the evaluation of power and sample size in superiority clinical trials with co-primary endpoints. We first discuss a simple case of a superiority trial comparing two interventions without an interim analyses. We then discuss extensions to include interim analyses, three-arm trials, and non-inferiority trials.

email: hamasakt@medstat.med.osaka-u.ac.jp

89. RECENT DEVELOPMENTS IN CHANGE POINT SEGMENTATION: FROM BIOPHYSICS TO GENETICS

HIGH THROUGHPUT ANALYSIS OF FLOW CYTOMETRY DATA WITH THE EARTH MOVER DISTANCE

Guenther Walther*, Stanford University
Noah Zimmermann, Stanford University

Changes in frequency and/or biomarker expression in small subsets of peripheral blood cells provide key diagnostics for disease presence, status and prognosis. At present, flow cytometry instruments that measure the joint expression of up to 20 markers in large numbers of individual cells are used to measure surface and internal marker expression. This technology is routinely used to inform therapeutic decision-making. Nevertheless, quantitative methods for comparing data between samples are sorely lacking. Based on the Earth Movers Distance, we describe novel computational methods that provide reliable indices of change in subset representation and/or marker expression by individual subsets of cells. We show that these methods are easily applied and readily interpreted.

e-mail: gwalther@stanford.edu

SIMULTANEOUS MULTISCALE CHANGE-POINT INFERENCE IN EXPONENTIAL FAMILIES: SHARP DETECTION RATES, CONFIDENCE BANDS, ALGORITHMS AND APPLICATIONS

Axel Munk* Goettingen University and max Planck Institute for Biophysical Chemistry
Klaus Frick, Goettingen University
Hannes Sieling, Goettingen University
Rebecca von der Heide, Goettingen University

We aim for estimating the number and locations of change-points of a piecewise constant regression function by minimizing the number of change-points over the acceptance region of a multiscale test. Deviation bounds for the estimated number of change points as well as convergence rates for the change-point locations close to the sampling rate 1/n are proven. We further derive the asymptotic distribution of the test statistic which can be used to construct asymptotically honest confidence bands for the regression function. We show how dynamic programming techniques can be employed for efficient computation of estimators and confidence regions and compare our method with state-of-the-art change-point detection methods in the recent literature. Finally, the performance of the proposed multiscale approach is illustrated, when applied to cutting-edge applications such as genetic engineering and photoemission spectroscopy.

e-mail: munk@math.uni-goettingen.de

CHANGE POINT SEGMENTATION FOR TIME DYNAMIC VOLTAGE DEPENDENT ION CHANNEL RECORDINGS

Rebecca von der Heide, Georgia Augusta University of Goettingen
Thomas Hotz*, Ilmenau University of Technology
Hannes Sieling, Georgia Augusta University of Goettingen
Claudia Steinem, Georgia Augusta University of Goettingen
Ole Schuette, Georgia Augusta University of Goettingen
Ulf Diederichsen, Georgia Augusta University of Goettingen
Tatjana Polupanow, Georgia Augusta University of Goettingen
Katarzyna Wasilczuk, Georgia Augusta University of Goettingen
Axel Munk, Georgia Augusta University of Goettingen

The characterization and reconstruction of ion channel functionalities is an important issue to understand cellular processes. For this purpose we suggest a new change-point detection technique to analyze current traces of ion channels. The underlying ion channel signal is assumed to be piecewise constant. For many membrane gating proteins the ion channel signals are in the order of picoampere resulting in a low signal-to-noise ratio. To overcome this burden we suggest a locally adaptive approach which is built on a multiscale statistic and only assumes a locally constant signal. Furthermore, we generalize our approach to time-varying exogenous variables to analyze channels with time-dependent dynamics like human voltage dependent anion channels. The performance of our method is demonstrated by simulations and applied to various ion channel data sets. To this end a graphical user interface (stepR) for process evaluation of ion channel recordings has been developed.

e-mail: rvonder1@uni-goettingen.de

STEPWISE SIGNAL EXTRACTION VIA MARGINAL LIKELIHOOD

Chao Du*, Stanford University
Samuel C. Kou, Harvard University

We propose a new method to estimate the number and locations of change-points in stepwise signal. Our approach treats each possible set of change-points as an individual model and uses marginal likelihood as the model section tool. Under an independence assumption of the parameters between successive change-points, the computational complexity of this approach is at most quadratic in the number of observations using a dynamic programming algorithm. The asymptotic properties of the marginal likelihood are studied. This paper further discusses the impact of the prior on the estimation and provide guidelines in choosing the prior. Detailed simulation study is carried out to compare the effectiveness of this method with other existing methods. We demonstrate this approach on DNA array CGH data and single molecule enzyme data. Our study shows that this method is capable of coping with a wide range of models and has appealing properties in applications.

e-mail: chaodu@stanford.edu

90. NEW CHALLENGES FOR NETWORK DATA AND GRAPHICAL MODELING

CONSISTENCY OF COMMUNITY DETECTION

Yunpeng Zhao, George Mason University
Elizaveta Levina, University of Michigan
Ji Zhu*, University of Michigan

Community detection is a fundamental problem in network analysis, with applications in many diverse areas. The stochastic block model is a common tool for model-based community detection, and asymptotic tools for checking consistency of community detection under the block model have been recently developed (Bickel and Chen, 2009). However, the block model is limited by its assumption that all nodes within a community are stochastically equivalent, and provides a poor fit to networks with hubs or highly varying node degrees within communities, which are common in practice. The degree-corrected block model (Karrer and Newman, 2010) was proposed to address this shortcoming, and allows variation in node degrees within a community while preserving the overall block model community structure. In this paper, we establish general theory for checking consistency of community detection under the degree-corrected block model, and compare several community detection criteria under both the standard and the degree-corrected block models.

e-mail: jizhu@umich.edu

SPARSE ESTIMATION OF CONDITIONAL GRAPHICAL MODELS WITH APPLICATION TO GENE NETWORKS

Bing Li*, The Pennsylvania State University
Hyonho Chun, Purdue University
Hongyu Zhao, Yale University

In many applications the graph structure in a network arises from two sources: intrinsic connections and connections due to external effects. We introduce a sparse estimation procedure for graphical models that is capable of isolating the intrinsic connections by removing the external effects. Technically, this is formulated as a conditional graphical model, in which the external effects are modeled as predictors, and the graph is determined by the conditional precision matrix. We introduce two sparse estimators of this matrix using reproducing kernel Hilbert space combined with lasso and adaptive lasso. We establish the sparsity, variable selection consistency, oracle property, and the asymptotic distributions of the proposed estimators. We also develop their convergence rate when the dimension of the conditional precision matrix goes to infinity. The methods are compared with sparse estimators for unconditional graphical models, and with the constrained maximum likelihood estimate that assumes a known graph structure. The methods are applied to a genetic data set to construct a gene network conditioning on single-nucleotide polymorphisms.

e-mail: bing@stat.psu.edu

MODEL-BASED CLUSTERING OF LARGE NETWORKS

Duy Q. Vu, University of Melbourne
David R Hunter*, The Pennsylvania State University
Michael Schweinberger, The Pennsylvania State University

We describe a network clustering framework, based on finite mixture models, that can be applied to discrete-valued networks with hundreds of thousands of nodes and billions of edge variables. Relative to other recent model-based clustering work for networks, we introduce a more flexible modeling framework, improve the variational-approximation estimation algorithm, discuss and implement standard error estimation via a parametric bootstrap approach, and apply these methods to much larger datasets than those seen elsewhere in

the literature. The more flexible modeling framework is achieved through introducing novel parameterizations of the model, giving varying degrees of parsimony, using exponential family models whose structure may be exploited in various theoretical and algorithmic ways. The algorithms, which we show how to adapt to the more complicated optimization requirements introduced by the constraints imposed by the novel parameterizations we propose, are based on variational generalized EM algorithms, where the E-steps are augmented by a minorization-maximization (MM) idea. The bootstrapped standard error estimates are based on an efficient Monte Carlo network simulation idea. Last, we demonstrate the usefulness of the model-based clustering framework by applying it to a discrete-valued network with more than 131,000 nodes and 17 billion edge variables.

e-mail: dhunter@stat.psu.edu

MAXIMUM LIKELIHOOD ESTIMATION OF A DIRECTED ACYCLIC GAUSSIAN GRAPH

Yiping Yuan, University of Minnesota
Xiaotong Shen*, University of Minnesota
Wei Pan, University of Minnesota

Directed acyclic graphs have been widely used to describe causal relations among interacting units. Estimation of a directed acyclic graph presents a great challenge without prior knowledge about the order of interacting units, where the number of enumeration of potential directions grows super-exponentially. A traditional method usually estimates directions locally and sequentially, and hence results in biased estimation. In this paper, we propose a global approach to determine all directions simultaneously, through constrained maximum likelihood with nonconvex constraints reinforcing a directed acyclic graph requirement. Computationally, we propose an efficient algorithm based on a projection-based accelerated gradient method and difference convex programming for approximating nonconvex constrained sets. Numerically, we demonstrate that the method leads to accurate parameter estimation, in parameter estimation as well as identifying graphical structures. Moreover, an application to gene network analysis will be described.

e-mail: xshen@umn.edu

91. BAYESIAN ANALYSIS OF HIGH DIMENSIONAL DATA

A MULTIVARIATE CAR MODEL FOR PRE-SURGICAL PLANNING WITH fMRI

Zhuqing Liu*, University of Michigan
Veronica J. Berrocal, University of Michigan
Timothy D. Johnson, University of Michigan

There is increasing interest in functional magnetic resonance imaging (fMRI) for pre-surgical planning. In a standard fMRI analysis, strict false positive control is desired. For pre-surgical planning, false negatives are of greater concern. In this talk, we present a multivariate intrinsic conditional autoregressive (MCAR) model and a new loss function that are designed to 1) leverage correlation between multiple fMRI contrast images within a single subject and 2) allow for asymmetric treatment of false positives and false negatives. Our model is a hierarchical model with an intrinsic MCAR prior. This model is used to incorporate information from multiple fMRI contrasts within the same subject, allowing simultaneous smoothing of the multiple contrast images. We apply the proposed model to a single subject's pre-surgical fMRI data to assess peri-tumoral brain activation. Our loss function treats false positives and false negatives asymmetrically allowing stricter control of false negatives. Through simulation studies we compare results from our MCAR model with results from (standard) univariate CAR models---that model the fMRI contrast images independently. Our model is further compared with a Bayesian non-parametric Potts model previously proposed (Johnson et al., 2011).

e-mail: zhuqingl@umich.edu

MODELING FUNCTIONAL CONNECTIVITY IN THE HUMAN BRAIN WITH INCORPORATION OF STRUCTURAL CONNECTIVITY

Wenqiong Xue*, Emory University
DuBois Bowman, Emory University

Recent innovations in neuroimaging technology have provided opportunities for researchers to investigate connectivity in the human brain by examining the anatomical circuitry as well as functional relationships between brain regions. Existing statistical approaches for connectivity generally examine resting-state or task-related functional connectivity (FC) between brain regions or separately examine structural linkages. We present a unified Bayesian framework for analyzing FC utilizing the knowledge of associated structural connections, which extends an approach by Patel et al. (2006a) that considers only functional data. Our FC measure rests upon assessments of functional coherence between regional brain activity identified from functional magnetic resonance imaging (fMRI) data. Our structural connectivity (SC) information is drawn from diffusion tensor imaging (DTI) data, which is used to quantify probabilities of SC between brain regions. We formulate a prior distribution for FC that depends upon the probability of SC between brain regions, with this dependence adhering to structure-function links revealed by our fMRI and DTI data. We further characterize the functional hierarchy of functionally connected brain regions by defining an ascendancy measure that compares the marginal probabilities of elevated activity between regions.

e-mail: wxue@emory.edu

A BAYESIAN SPATIAL POSITIVE-DEFINITE MATRIX REGRESSION MODEL FOR DIFFUSION TENSOR IMAGING

Jian Kang*, Emory University

There has been a growing interest in using diffusion tensor imaging (DTI) to provide information about anatomical connectivity in the brain. At each voxel, the DTI technique measures a diffusion tensor, i.e. a 3×3 symmetric positive-definite (SPD) matrix, to track the effective diffusion of water molecules. Motivated by this problem, Zhu et al (2009) developed a very first semiparametric regression model with SPD matrices as responses in a Remannian manifold and covariates in Euclidean space. However, this pioneer model ignores the spatial dependence in SPD matrices between the voxels, which is critical for the analysis of the DTI data. To address this limitation, in this work, we propose a nonparametric Bayesian spatial regression model for SPD matrices. We jointly model the three eigenvalue-eigenvector pairs of diffusion tensors over space using multiple Gaussian processes. The proposed model provides a framework to characterize the spatial correlation between diffusion tensors. We develop a Metropolis adjusted Langevin algorithm based on circulant embedding of the covariance matrix for efficient posterior computation. We illustrate the proposed methods on simulation studies and a diffusion tensor study of major depressive disorder.

e-mail: jian.kang@emory.edu

BAYESIAN SQUASHED REGRESSION

Rajarshi Guhaniyogi*, Duke University
David B. Dunson, Duke University

As an alternative to shrinkage priors in large p, small n problem, we propose "squashing" the high dimensional predictors to lower dimensions for estimation and inferential purpose. As opposed to shrinkage priors, an exact posterior distribution of parameters are available from the proposed model, avoiding convergence issues due to the model fitting by MCMC algorithm. Bayesian model averaging techniques are implemented to have better predictive performance for the proposed model. The proposed model also found to enjoy attractive theoretical properties, e.g. near parametric convergence rate for the predictive density.

e-mail: rg124@stat.duke.edu

GENERALIZED BAYESIAN INFINITE FACTOR MODELS

Kassie Fronczyk*, University of Texas MD Anderson Cancer Center and Rice University
Michele Guindani, University of Texas MD Anderson Cancer Center
Marina Vannucci, Rice University

Experiments generating high dimensional data are becoming more prevalent throughout the literature, with examples ranging from genomics and biology to imaging. A widely used approach to analyze this type of data is factor analysis, where the aim is to explain observations with a linear projection of independent hidden factors. Given a number of latent factors and loadings are random variables, traditional models assume some sort of isotropic or diagonal error covariance structure, gaining insight for correlations only across the rows of the data. We extend this idea to a more general setting, where interest lies in correlations of the rows and columns of the data and the number of factors is treated as an unknown parameter. We explore a fully Bayesian approach to obtain inference on the latent factors and loadings, as well as for the latent dimension of the data.

e-mail: kf8@rice.edu

A BAYESIAN MIXTURE MODEL FOR GENE NETWORK SELECTION

Yize Zhao*, Emory University

It is very challenging to select informative features from tens of thousands of measured features in high-throughput data analysis. Recently, several parametric/regression models have been developed utilizing the gene network information to select genes or pathways strongly associated with a clinical/biological outcome. Alternatively, in this paper, we propose a nonparametric Bayesian model for gene selection incorporating network information based on large scale statistics. In addition to identifying genes that have a strong association with a clinical outcome, our model can select genes with particular expressional behavior in which case the regression models are not directly applicable. We show that our proposed model is equivalent to an infinity mixture model for which we develop a posterior computation algorithm based on Markov chain Monte Carlo (MCMC) methods. We also propose two fast computing algorithms that approximate the posterior simulation with good accuracy of the gene selection but relatively low computational cost. We illustrate our methods on simulation studies and the analysis of Spellman yeast cell cycle microarray data.

e-mail: yize.zhao@emory.edu

BAYES MULTIPLE DECISION FUNCTIONS IN CLASSIFICATION

Wensong Wu*, Florida International University
Edsel A. Pena, University of South Carolina

In this presentation we consider a two-class classification problem, where the goal is to predict the class membership of M units based on the values of high-dimensional predictor variables as well as both the values of the predictor variables and the class membership of other N independent units. We consider a Bayesian and decision-theoretic framework, and develop a general form of Bayes multiple decision function (BMDF) with respect to a class of cost-weighted loss functions. In particular, the loss function pairs such as the proportions of false positives and false negatives, and (1-sensitivity) and (1-specificity), are considered, and the cost weights are pre-specified. An efficient algorithm of finding the BMDF is provided based upon posterior expectations. The result is applicable to general classification models, but particular generalized linear regression models are investigated, where the predictor variables and the link functions are to be chosen from a finite class. The results will be illustrated via simulations and on a Lupus diagnose dataset.

e-mail: wenswu@fiu.edu

92. MISSING DATA

A CLASS OF TESTS FOR MISSING COMPLETELY AT RANDOM

Gong Tang*, University of Pittsburgh

We consider regression analysis of data with nonresponse. When the nonresponse is missing at random, the ignorable likelihood method yields valid inference. However, the assumption of missing at random is not testable in general. A stronger assumption, missing completely at random (MCAR), is testable. Likelihood ratio tests have been discussed in the context of multivariate data with missing values but these tests require the specification of the joint distribution of all variables (Little, 1988). Subsequently Chen & Little (1999), and Qu & Song (2002) proposed a Wald-type test and a score test for generalized estimating equations with using the same fact that all sub-patterns share the same model parameters under MCAR. For regression analysis of data with nonresponse, Tang, Little and Raghunathan proposed a pseudolikelihood estimate without specifying the missing-data mechanism for a class of nonignorable mechanisms. In the line of this pseudolikelihood method, here we propose a class of Wald-type tests for MCAR by comparing the ignorable likelihood estimate and a class of pseudolikelihood estimates of regression parameters and evaluate their performance via simulation studies.

e-mail: got1@pitt.edu

ESTIMATION IN LONGITUDINAL STUDIES WITH NONIGNORABLE DROPOUT

Jun Shao, University of Wisconsin, Madison
Jiwei Zhao*, Yale University

A sampled subject with repeated measurements often drops out prior to the study end. Data observed from such a subject is longitudinal with monotone missing. If dropout at a time point t is only related to past observed data from the response variable, then it is ignorable and statistical methods are well developed. When dropout is related to the possibly missing response at t even after conditioning on all past observed data, it is nonignorable and statistical analysis is difficult. Without any further assumption, unknown parameters may not be identifiable when dropout is nonignorable. We develop a semiparametric pseudo likelihood method that produces consistent and asymptotically normal estimators under nonignorable dropout with the assumption that there exists a dropout instrument, a covariate related to the response variable but not related to the dropout conditioned on the response and other covariates. Our main effort is to derive easy-to-compute consistent estimators of the asymptotic covariance matrices for assessing variability or inference. For illustration, we present an example using the HIV-CD4 data and some simulation results.

e-mail: jiwei.zhao@yale.edu

SEMIPARAMETRICALLY EFFICIENT ESTIMATION IN LONGITUDINAL DATA ANALYSIS WITH DROPOUTS

Peisong Han*, University of Michigan
Peter X. K. Song, University of Michigan
Lu Wang, University of Michigan

Longitudinal data with dropouts are commonly encountered in practical studies, especially in clinical trials. Methods based on weighted estimating functions, including the inverse probability weighted (IPW) generalized estimating equations and the augmented inverse probability weighted (AIPW) generalized estimating equations, are widely used in analyzing longitudinal data with dropouts. However, these methods hardly yield efficient estimation, even if the within-subject correlation is correctly modeled. This is because the estimating function that is optimal with fully observed data would lose its estimation optimality when missing data exist. In this paper we propose an empirical-likelihood-based method. Our method does not need to construct any estimating functions, hence avoids the modeling of the variance-covariance of longitudinal outcomes. Assuming that the dropouts follow the missing at random mechanism, we show that our proposed method produces a doubly robust and locally efficient estimator of the regression coefficients.

e-mail: peisong@umich.edu

WEIGHTED ESTIMATING EQUATIONS FOR SEMIPARAMETRIC TRANSFORMATION MODELS WITH MISSING COVARIATES

Yang Ning*, University of Waterloo
Grace Yi, University of Waterloo
Nancy Reid, University of Toronto

In survival analysis, covariate measurements often contain missing observations. Ignoring missingness in covariates can result in biased results. To conduct valid inferences, properly adjusting for missingness effects is usually necessary. We propose a weighted estimating equation approach to handle missing covariates under semiparametric transformation models for right censored data. The weights are determined by the missingness probabilities, which are modeled both parametrically and nonparametrically. To improve efficiency, the weighted estimating equations are augmented by another sets of unbiased estimating equations. The proposed estimators are shown to be consistent and asymptotically normal. Finite sample performance of the estimators is evaluated by empirical studies.

e-mail: yning@jhsph.edu

HANDLING DATA WITH THREE TYPES OF MISSING VALUES

Jennifer Boyko*, University of Connecticut

Incomplete data is a common obstacle to the analysis of data in a variety of fields. Values in a data set can be missing for several different reasons including failure to answer a survey question, dropout, planned missing values, intermittent missed measurements, latent variables, and equipment malfunction. In fact, many studies will have more than just one type of missing value. Appropriately handling missing values is critical in the inference for a parameter of interest. Many methods of handling missing values inappropriately fail to account for the uncertainty due to missing values. This failure to account for uncertainty can lead to biased estimates and over-confident inferences. One area which is still unexplored is the situation where there are three types of missing values in a study. This complication arises often in studies involved with cognitive functioning. These studies tend to have large amounts of missing values of several different types. I am proposing the development of a three stage multiple imputation approach which would be beneficial in analyzing these types of studies. Three stage multiple imputation would also extend the benefits of standard multiple imputation and two stage multiple imputation, namely the quantification of the variability attributable to each type of missing value and the flexibility for greater specificity regarding data analysis.

e-mail: jen.boyko@gmail.com

A BAYESIAN SENSITIVITY ANALYSIS MODEL FOR DIAGNOSTIC ACCURACY TESTS WITH MISSING DATA

Chenguang Wang*, Johns Hopkins University
Qin Li, U.S. Food and Drug Administration
Gene Pennello, U.S. Food and Drug Administration

When comparing the accuracy of a new diagnostic medical device to a predicate device on the market, missing data, including missing reference standard and missing device test results, are commonly encountered. The missing data challenge could be two-folded in the context of diagnostic accuracy test. First, as in general cases, a sensitivity analysis model is needed to account for the uncertainty of the missing data mechanisms. Second, the often made assumption that the two tests of the new and the predicate devices are conditionally independent cannot be verified with missing data and sensitivity analysis with respect to such an independence assumption is required. In this paper, we propose an integrated Bayesian framework to address both two issues. We expect our proposal to be widely used for the regulatory's diagnostic device approval decision making process.

e-mail: cwang68@jhmi.edu

MULTIPLE IMPUTATION MODEL DIAGNOSTICS

Irina Bondarenko*, University of Michigan
Trivellore Raghunathan, University of Michigan

Multiple imputation is a popular approach for analyzing incomplete data. Software packages have become widely available for imputing the missing values. However, diagnostic tools to check the validity of imputations are limited. We propose a set of diagnostic tools that compares certain conditional distributions of the observed and imputed values to assess if imputations are reasonable under the Missing At Random (MAR) assumption. Proposed method does not require knowledge of the exact model used for creating imputations. Offered diagnostics are useful to identify whether a variable, important for the analyst, has been omitted from the imputation process, and to assess a need for more elaborate imputation model that includes interactions, or non-linear terms. The method is illustrated using a dataset with large number of variables from different distribution families. Performance of the method is assessed using simulated datasets.

e-mail: ibond@umich.edu

93. SEMIPARAMETRIC AND NONPARAMETRIC METHODS FOR SURVIVAL ANALYSIS

BAYESIAN PARTIAL LINEAR MODEL FOR SKEWED LONGITUDINAL DATA

Yuanyuan Tang*, Florida State University
Debajyoti Sinha, Florida State University
Debdeep Pati, Florida State University
Stuart Lipsitz, Brigham and Women's Hospital

For longitudinal studies with heavily skewed continuous response, statistical model and methods focusing on mean response are not appropriate. In this paper, we present a partial linear model of median regression function of skewed longitudinal response. We develop a semiparametric Bayesian estimation procedure using an appropriate Dirichlet process mixture prior for the skewed error distribution. We provide justifications for using our methods including theoretical investigation of the support of the prior, asymptotic properties of the posterior and also simulation studies of finite sample properties. Ease of implementation and advantages of our model and method compared to existing methods are illustrated via analysis of a cardiotoxicity study of children of HIV infected mother.

e-mail: ytang@stat.fsu.edu

ON ESTIMATION OF GENERALIZED TRANSFORMATION MODEL WITH LENGTH-BIASED RIGHT-CENSORED DATA

Mu Zhao*, Northwestern University
Hongmei Jiang, Northwestern University
Yong Zhou, Academy of Mathematics and Systems Science, Chinese Academy of Sciences

Length-biased sampling, in which the observed samples are not randomly drawn from the population of interest but with probability proportional to their length, has been well recognized in cancer screening studies, epidemiological, economics and industrial reliability. In this paper, we propose to use general transformation models for regression analysis with length-biased data. With unknown link function and unknown error distribution, the transformation models are broad enough to cover accelerated failure time model, Cox proportional model, proportional odds model, additive hazard model and Box and Cox transformation model. We propose monotone rank estimators (MRE) to estimate the regression parameters. Without the need of estimating transformation function and unknown error distribution, the proposed procedure is numerically easy to implement with the Nelder-Mead simplex algorithm. Consistency and normality of the proposed estimates are established. Some simulations and a real data example are also presented to illustrate our results.

e-mail: mu.zhao@northwestern.edu

TIME-VARYING COPULA MODELS FOR LONGITUDINAL DATA

Esra Kurum, Istanbul Medeniyet University
John Hughes*, University of Minnesota
Runze Li, The Pennsylvania State University

We propose a joint modeling framework for mixed longitudinal responses of practically any type and dimension. Our approach permits all model parameters to vary with time, and thus will enable researchers to reveal dynamic response-predictor relationships and response-response associations. We call the new class of models timecop because we model dependence using a time-varying copula. We develop a one-step estimation procedure for the parameter vector, and also describe how to estimate standard errors. We investigate the finite sample performance of our procedure via a Monte Carlo simulation study, and illustrate the applicability of our approach by estimating the time-varying association between binary and continuous responses from the Women's Interagency HIV Study. Our methods are implemented in an easy-to-use software package, freely available from the Comprehensive R Archive Network.

e-mail: hughesj@umn.edu

REGRESSION ANALYSIS OF CURRENT STATUS DATA USING THE EM ALGORITHM

Christopher S. McMahan*, Clemson University
Lianming Wang, University of South Carolina
Joshua M. Tebbs, University of South Carolina

We propose new expectation-maximization algorithms to analyze current status data under two popular semiparametric regression models: the proportional hazards (PH) model and the proportional odds (PO) model. Monotone splines are used to model the baseline cumulative hazard function in the PH model and the baseline odds function in the PO model. The proposed algorithms are derived by exploiting a data augmentation based on Poisson latent variables. Unlike previous regression work with current status data, our PH and PO model fitting methods are fast, flexible, easy to implement, and provide variance estimates in closed form. These techniques are evaluated using simulation and are illustrated using uterine fibroid data from a prospective cohort study on early pregnancy.

e-mail: mcmaha2@clemson.edu

QUANTILE REGRESSION FOR LONGITUDINAL STUDIES WITH MISSING AND LEFT CENSORED MEASUREMENTS

Xiaoyan Sun*, Emory University
Limin Peng, Emory University
Amita K. Manatunga, Emory University
Robert H. Lyles, Emory University
Michele Marcus, Emory University

Epidemiological follow up studies often present various challenges that can complicate statistical analysis. For example, in the Michigan polybrominated biphenyl (PBB) study, the longitudinal serum PBB measurement is subject to left censoring due to laboratory assay detection limit while the data are highly suggestive of a subject follow-up pattern depending on initial observed PBB concentrations. In this work, we consider quantile regression modeling for the data from such longitudinal studies. We adopt an appropriate censored quantile regression technique to handle left censoring and employ the idea of inverse probability weighting to tackle the issue associated with informative intermittent missing mechanism. We evaluate our method by simulation studies. The proposed method is applied to the Michigan PBB study to investigate the PBB decay profile.

e-mail: xsun33@emory.edu

QUANTILE REGRESSION OF SEMIPARAMETRIC TIME VARYING COEFFICIENT MODEL WITH LONGITUDINAL DATA

Xuerong Chen*, University of Missouri, Columbia
Jianguo Sun, University of Missouri, Columbia

Longitudinal data often arises in medical follow-up studies and economic research. Semiparametric models are often considered for analyzing longitudinal data. In this paper, we propose a semiparametric time varying coefficient quantile regression model for analysis of longitudinal data in the presence of informative observation times. The time varying coefficients are approximated by basis function approximations. The asymptotic properties of the proposed estimators are established for the time varying coefficients as well as for the constant coefficients. The small sample properties of the proposed procedure are investigated in a Monte Carlo study and a real data example illustrates the application of the method in practice.

e-mail: chenxr522@yahoo.cn

LONGITUDINAL ANALYSIS OF THE LEUKOCYTE AND CYTOKINE FLUCTUATIONS AFTER STEM CELL TRANSPLANTATION USING VARYING COEFFICIENT MODELS

Xin Tian*, National Heart, Lung and Blood Institute, National Institutes of Health

Allogeneic hematopoietic stem cell transplantation for hematologic malignancy is associated with profound changes in levels of leukocytes and various cytokines. It was unclear the relationship of the production and fluctuations of cytokines in response to the cytopenia and lymphocyte recovery in the peri- and immediate post-transplant period. We propose to use mixed -effects varying-coefficient model to model the longitudinal variation and correlation of leukocytes and cytokines. Flexible nonparametric regression splines are used for inference.

e-mail: tianx@nhlbi.nih.gov

94. MEASUREMENT ERROR

THRESHOLD-DEPENDENT PROPORTIONAL HAZARDS MODEL FOR CURRENT STATUS DATA WITH BIOMARKER SUBJECT TO MEASUREMENT ERROR

Noorie Hyun*, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
David J. Couper, University of North Carolina, Chapel Hill
James S. Pankow, University of Minnesota

In many medical studies, the time to a disease event is determined by the value of some biomarker crossing a specified threshold. Current status data arise when the biomarker value at a visit time is above or below the corresponding threshold. However, assuming a fixed threshold for all subjects may not be appropriate for some biomarkers. Medical researchers have showed that thresholds can vary across populations or from person to person. Furthermore, a biomarker is usually subject to measurement error. In the presence of these two challenging issues, existing methods for analyzing current status data are no longer applicable. In this paper, we propose a semiparametric method based on the Cox regression model depending on threshold values to account for measurement error in the biomarker. We estimate the model parameters using the nonparametric maximum likelihood approach and implement computation via the EM algorithm. We show consistency and semiparametric efficiency of the regression parameter estimator and estimate asymptotic variance. The method is illustrated through an application to data from a diabetes study.

e-mail: nrhyun@live.unc.edu

PROPORTIONAL HAZARDS MODEL WITH FUNCTIONAL COVARIATE MEASUREMENT ERROR AND INSTRUMENTAL VARIABLES

Xiao Song*, University of Georgia
Ching-Yun Wang, Fred Hutchinson Cancer Research Center

In biomedical studies, covariates with measurement error may occur in survival data. Existing approaches mostly require certain replications on the error-contaminated covariates, which may not be available in the data. In this paper, we develop a simple nonparametric correction approach for the proportional hazards model using measurements on instrumental variables observed in a subset of the sample. The instrumental variable is related to the covariates through a general nonparametric model, and no distributional assumptions are placed on the error and the underlying true covariates. We further propose a novel generalized methods of moments nonparametric correction estimator to improve the efficiency over the simple correction approach. The efficiency gain can be substantial when the calibration subsample is small compared to the whole sample. The estimators are shown to be consistent and asymptotically normal. Performance of the estimators is evaluated via simulation studies and by an application to data from an HIV clinical trial.

e-mail: xsong@uga.edu

DISTANCE AND GRAVITY: MODELING CONDITIONAL DISTRIBUTIONS OF HEAPED SELF-REPORTED COUNT DATA

Sandra D. Griffith*, Cleveland Clinic
Saul Shiffman, University of Pittsburgh
Daniel F. Heitjan, University of Pennsylvania

Self-reported daily cigarette counts typically exhibit measurement error, often manifesting as a preponderance of round numbers. Heaping, a form of measurement error that occurs when quantities are reported with varying levels of precision, offers one explanation. A doubly-coded data set with both a conventional retrospective recall measurement (timeline followback) and an instantaneous measurement with a smooth distribution (ecological momentary assessment), allows us to model the conditional distribution of a self-reported count given the underlying true count. Our model incorporates notions from cognitive psychology to conceptualize a subject's selection of a self-reported count as a function of both its distance from the true value and an intrinsic attractiveness of the reported numeral, which we denote its gravity. We develop a flexible framework for parameterizing the model, allowing gravities based on the roundness of numerals or data-driven gravities based on empirical frequencies. When applied to the motivating cigarette consumption data, the frequency-based gravity model produced the better fit. This method holds potential for application to a wide range of self-reported count data.

e-mail: griffis5@ccf.org

PATHWAY ANALYSIS OF GENE-ENVIRONMENT INTERACTIONS IN THE PRESENCE OF MEASUREMENT ERROR IN THE ENVIRONMENTAL EXPOSURE

Stacey E. Alexeeff*, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

Many complex disease processes are thought to be influenced by a number of genetic and environmental factors. Pathway analysis is a growing area of methodological research, where the objective is to identify a set of genetic and environmental risk factors that can explain a meaningful proportion of disease susceptibility. Since genes and environmental exposures on the same biological pathway may interact functionally, there is growing scientific interest to study these sets of factors in pathway analysis. A score test can account for the correlation among the covariates in a test for pathway effect. Genes on the same pathway are expected to be correlated and environmental exposures may also be correlated with genetic covariates. We consider the impact of measurement error in the environmental exposure in a pathway test for the effect of a gene-environment interaction. Measurement error in the environmental exposure impacts the gene-environment interaction terms. We investigate how this error propogates through a linear health effects model, biases the coefficients and ultimately affects the score test for pathway effect.

e-mail: salexeeff@fas.harvard.edu

VARIABLE SELECTION FOR MULTIVARIATE REGRESSION CALIBRATION WITH ERROR-PRONE AND ERROR-FREE COVARIATES

Xiaomei Liao*, Harvard School of Public Health
Kathryn Fitzgerald, Harvard School of Public Health
Donna Spiegelman, Harvard School of Public Health

Regression calibration is a popular method for correcting for bias in effect estimates when a disease risk factor is measured with error. However, the development of such methods has thus far been focused on unbiased estimation and inference for the primary exposure or exposures of interest, and not on finer aspects of model building and variable selection. Adjusting for measurement error using the regression calibration method evokes several questions concerning valid model construction in the presence of covariates. For instance, are the standard regression calibration adjustments valid when a covariate that is associated only with the measurement error process and not the outcome itself is included in the primary regression model? Does the inclusion of such a variable in the primary regression model induce extraneous variation in the resulting estimator? Clear answers to these questions would provide valuable insight and improve estimation of exposure disease associations measured with error. In the paper, we address these questions analytically and develop extended regression calibration estimators as needed based on assumptions about the underlying association between disease and covariates, for both linear and logistic regression models. The methods are applied to data from the Nurses' Health Study.

e-mail: stxia@channing.harvard.edu

DISK DIFFUSION BREAKPOINT DETERMINATION USING A BAYESIAN NONPARAMETRIC VARIATION OF THE ERRORS-IN-VARIABLES MODEL

Glen DePalma*, Purdue University
Bruce A. Craig, Purdue University

Drug dilution (MIC) and disk diffusion (DIA) are the two antimicrobial susceptibility tests used by hospitals and clinics to determine an unknown pathogen's susceptibility to various antibiotics. Both tests classify the pathogen as either being susceptible, indeterminant, or resistant to a drug. Since only one of these tests will typically be used in practice, it is imperative to have the two tests calibrated. The MIC test deals with concentrations of a drug so its classification breakpoints are based primarily on a drug's pharmacokinetics and pharmacodynamics. The DIA test, on the other hand, does not and therefore its breakpoints are determined by minimizing classification discrepancies between pairs of MIC and DIA test results. It has been shown that this minimization procedure does not adequately account for the inherent variability and unique properties of each test and as a result, produces biased and imprecise breakpoints. In this paper we present a hierarchical errors-in-variables model that explicitly accounts for these various factors of uncertainty and uses estimated probabilities from the model to determine appropriate breakpoints. We show through a simulation study that this method leads to more accurate and precise results.

e-mail: gdepalma@purdue.edu

95. GRAPHICAL MODELS

A BAYESIAN GRAPHICAL MODEL FOR INTEGRATIVE ANALYSIS OF TCGA DATA

Yanxun Xu*, Rice University and University of Texas MD Anderson Cancer Center
Jie Zhang, University of Texas MD Anderson Cancer Center
Yuan Yuan, University of Texas MD Anderson Cancer Center
Riten Mitra, University of Texas, Austin
Peter Muller, University of Texas, Austin
Yuan Ji, NorthShore University Health System

We integrate three TCGA data sets including measurements on matched DNA copy numbers (C), DNA methylation (M), and mRNA expression (E) over 500+ ovarian cancer samples. The integrative analysis is based on a Bayesian graphical model treating the three types of measurements as three vertices in a network. The graph is used as a convenient way to parameterize and display the dependence structure. Edges connecting vertices infer specific types of regulatory relationships. For example, an edge between M and E and a lack of edge between C and E implies methylation-controlled transcription, which is robust to copy number changes. In other words, the mRNA expression is sensitive to methylational variation but not copy number variation. We apply the graphical model to each of the genes in the TCGA data independently and provide a comprehensive list of inferred profiles. Examples are provided based on simulated data as well.

e-mail: yanxun.xu@rice.edu

BAYESIAN INFERENCE OF MULTIPLE GAUSSIAN GRAPHICAL MODELS

Christine B. Peterson*, Rice University
Francesco C. Stingo, University of Texas MD Anderson Cancer Center
Marina Vannucci, Rice University

We propose a Bayesian method for inferring multiple Gaussian graphical models that are believed to share common features, but may differ in scientifically important respects. In our model, we place a Markov Random Field prior on the network structure for each sample group that both encourages similar structure between related groups and accounts for reference networks established by previous research. This formulation improves the reliability of the estimated networks by allowing us to borrow strength across related sample groups and encouraging similarity to a known network. Applications include comparison of the cellular metabolic networks for control vs. disease groups, and the inference of protein-protein interaction networks for multiple cancer subtypes.

e-mail: cbpeterson@gmail.com

DIFFERENTIAL PATTERNS OF INTERACTION AND GAUSSIAN GRAPHICAL MODELS

Masanao Yajima*, University of California, Los Angeles
Donatello Telesca, University of California, Los Angeles
Yuan Ji, NorthShore University HealthSystem
Peter Muller, University of Texas, Austin

We propose a methodological framework to assess heterogeneous patterns of association amongst components of a random vector expressed as a Gaussian directed acyclic graph. The proposed framework is likely to be useful when primary interest focuses on potential contrasts characterizing the association structure between known subgroups of a given sample. We provide inferential frameworks as well as an efficient computational algorithm to fit such a model and illustrate its validity through a simulation in the supplementary material. We apply the model to Reverse Phase Protein Array data on Acute Myeloid Leukemia patients to show the contrast of association structure between refractory patients and relapsed patients.

e-mail: yajima@ucla.edu

PenPC: A TWO-STEP APPROACH TO ESTIMATE THE SKELETONS OF HIGH DIMENSIONAL DIRECTED ACYCLIC GRAPHS

Min Jin Ha*, University of North Carolina, Chapel Hill
Wei Sun, University of North Carolina, Chapel Hill
Jichun Xie, Temple University

Estimation of the skeleton of a directed acyclic graph (DAG) is of great importance for understanding the underlying DAG and causal effects can be assessed from the skeleton when the DAG is not identifiable. We propose a novel method named 'PenPC' to estimate the skeleton of a high-dimensional DAG by a two-step approach. We first estimate the non-zero entries of a concentration matrix using penalized regression, and then fix the difference between the concentration matrix and the skeleton by evaluating a set of conditional independence hypotheses. As illustrated by extensive simulations and real data studies, PenPC has significantly higher sensitivity and specificity than the standard-of-the-art method, the PC algorithm. We systematically study the asymptotic property of PenPC on high dimensional problem (the number of vertices p is in either polynomial or exponential scale of sample size n) of traditional random graph model where the number of connections of each vertex is limited and scale-free DAGs where one vertex may be connected to a large number of neighbors.

e-mail: mjha@live.unc.edu

JOINT ESTIMATION OF MULTIPLE DEPENDENT GAUSSIAN GRAPHICAL MODELS WITH APPLICATION TO TISSUE-SPECIFIC GENE EXPRESSION

Yuying Xie*, University of North Carolina, Chapel Hill
William Valdar, University of North Carolina, Chapel Hill
Yufeng Liu, University of North Carolina, Chapel Hill

Gaussian graphical models are widely used to represent conditional dependence among random variables. In this paper we propose a novel estimator for such models appropriate for data arising from several dependent graphical models. In this set- ting, existing methods that assume independence among graphs are not applicable. To estimate multiple dependent graphs, we decompose those graphical models into two layers: the systemic layer, which is the network shared among graphs and which induces across-graphs dependency, and the category-specific layer, which represents graph-specific variation. We propose a new graphical EM technique that jointly estimates the two layers of graphs aiming to learn the systemic network shared by graphs, as well as the category-specific network taking account the dependence of data. We establish the estimation consistency and selection sparsistency of the proposed estimator, and confirm the superior performance of the EM method over the naive one step method through simulations. Lastly, we apply our graphical EM technique to mouse genetic data and obtain biologically plausible results.

e-mail: xyy@email.unc.edu

GRAPHICAL NETWORK MODELS FOR MULTI-DIMENSIONAL NEUROCOGNITIVE PHENOTYPES OF PEDIATRIC DISORDERS

Vivian H. Shih*, University of California, Los Angeles
Catherine A. Sugar, University of California, Los Angeles

The rapidly emerging field of phenomics -- the study of dimensional patterns of deficits characterizing specific disorders -- plays a transformative role in fostering breakthroughs in neuropsychiatric research. Multi-dimensional relationships among neurocognitive constructs and intricate interactions between genes and behaviors appear not only within a specific disorder but also cut across current diagnostic boundaries. Traditional dimension reduction approaches such as principle component analysis and factor analysis generate new domains by collapsing across phenotypes before analysis, possibly losing substantial information. On the other hand, graphical network models constructively search for sparse relational structures of phenotypes within and across groups without the need of collapsing. This sparse covariance estimation technique provides a holistic view of the interconnectedness of phenotypic measures as well as specific hotspots within the underlying data structure. We

uncover vital phenotypes for childhood neuropsychiatric disorders (e.g., ADHD, autism, 22q11.2 deletion syndrome, and tic disorder) using the conventional estimation and modify the algorithm to reflect adjustments for other covariates and longitudinal patterns across time.

e-mail: vivianhshih@gmail.com

96. ADVANCES IN ROBUST ANALYSIS OF LONGITUDINAL DATA

NONPARAMETRIC RANDOM COEFFICIENT MODELS FOR LONGITUDINAL DATA ANALYSIS: ALGORITHMS; ROBUSTNESS, AND EFFICIENCY

John M. Neuhaus*, University of California, San Francisco
Charles E. McCulloch, University of California, San Francisco
Mary Lesperance, University of Victoria, Canada
Rabih Saab, University of Victoria, Canada

Generalized linear mixed models with random intercepts and slopes provide useful analyses of longitudinal data. Since little information exists to guide the choice of a parametric model for the distribution of random effects, several investigators have proposed approaches that leave this distribution unspecified, but these approaches have focussed on models with only random intercepts. In this talk we present an algorithm for fitting mixed effects models with a nonparametric joint distribution of slopes and intercepts. Using analytic and simulation studies, we compare the performance of this approach to that of fully parametric models with regard to bias and efficiency/ mean square error in settings with correct and incorrect specification of the joint distribution of random intercepts and slopes. Fits of the nonparametric and fully parametric mixed effects models to example data from longitudinal studies further illustrate the findings.

e-mail: john@biostat.ucsf.edu

ROBUST INFERENCE FOR MARGINAL LONGITUDINAL GENERALIZED LINEAR MODELS

Elvezio M. Ronchetti*, University of Geneva, Switzerland

Longitudinal models are commonly used for studying data collected on individuals repeatedly through time and classical statistical methods are readily available to carry out estimation and inference. However, in the presence of small deviations from the assumed model, these techniques can lead to biased estimates, p-values, and confidence intervals. Robust statistics deals with this problem and develops techniques that are not unduly influenced by such deviations. In this talk we first review several robust estimators for marginal longitudinal GLM which have been proposed in the literature together with the corresponding robust inferential procedures. Then we discuss robust variable selection procedures (including a generalized version of Mallows's Cp) and we examine their performance in the longitudinal setup. Finally, longitudinal data typically derive from medical or other large-scale studies where often large numbers of potential explanatory variables and hence even larger numbers of candidate models must be considered. In this case we discuss a cross-validation Markov Chain Monte Carlo procedure as a general variable selection tool which avoids the need to visit all candidate models. Inclusion of a "one-standard error" rule provides users with a collection of good models.

e-mail: Elvezio.Ronchetti@unige.ch

ROBUST ANALYSIS OF LONGITUDINAL DATA WITH INFORMATIVE DROP-OUTS

Sanjoy Sinha*, Carleton University
Abdus Sattar, Case Western Reserve University

In this talk, I will discuss a robust method for analyzing longitudinal data when there are informative drop-outs. This robust method is developed in the framework of weighted generalized estimating equations, and is useful for bounding the influence of potential outliers in the data when estimating the model parameters. The weights considered are inverse probabilities of responses, and are estimated robustly in the framework of a pseudo-likelihood. The empirical properties of the robust estimators will be discussed using simulations. An application of the proposed robust method will also be presented using real data from genetic and inflammatory markers of a sepsis study, which is a large cohort clinical study.

e-mail: sinha@math.carleton.ca

INFORMATIVE OBSERVATION TIMES IN LONGITUDINAL STUDIES

Kay-See Tan, University of Pennsylvania
Benjamin French*, University of Pennsylvania
Andrea B. Troxel, University of Pennsylvania

In longitudinal studies in medicine, subjects may be repeatedly observed according to a specified schedule, but they may also have additional visits for a variety of reasons. In many applications, these additional visits, and the times at which they occur, are informative in the sense that they are associated with the outcome of interest. An example is warfarin dosing and maintenance, in which subjects must be assessed regularly to ensure that their blood clotting activity is within the normal range. Subjects outside the normal range must return to the clinic at much closer intervals until they are once again within range. In this talk, I will review methods for addressing this problem, including inverse-intensity-weighted GEEs and joint modeling approaches, and present our recent extensions that address binary data and incorporate latent variables.

e-mail: atroxel@mail.med.upenn.edu

97. COMPLEX DESIGN AND ANALYTIC ISSUES IN GENETIC EPIDEMIOLOGIC STUDIES

USING FAMILY MEMBERS TO AUGMENT GENETIC CASE-CONTROL STUDIES OF A LIFE-THREATENING DISEASE

Lu Chen*, University of Pennsylvania School of Medicine
Clarice R. Weinberg, National Institute of Environmental Health Sciences, National Institutes of Health
Jinbo Chen*, University of Pennsylvania School of Medicine

Survival bias in case-control genetic association studies may arise due to an association between survival and genetic variants under study. It is difficult to adjust for the bias if no genetic data are available from deceased cases. We propose to incorporate genotype data from family members (such as offspring, spouses, or parents) of deceased cases into retrospective maximum likelihood analysis. Our method provides a partial data approach for correcting survival bias and for obtaining unbiased estimates of association parameters with di-allelic SNPs. This method faces an identifiability issue under a co-dominant model for both penetrance and survival given disease, so model simplifications are required. We derived closed-form maximum likelihood estimates for association parameters under the widely used log-additive and dominant association models. Our proposed method can improve both validity and study power by enabling inclusion of deceased cases, and we provide simulations to demonstrate achievable improvements in efficiency.

e-mail: jinboche@mail.med.upenn.edu

CASE-SIBLING STUDIES THAT ACKNOWLEDGE UNSTUDIED PARENTS AND PERMIT UNMATCHED INDIVIDUALS

Min Shi*, National Institute of Environmental Health Sciences, National Institutes of Health
David M. Umbach, National Institute of Environmental Health Sciences, National Institutes of Health
Clarice R. Weinberg, National Institute of Environmental Health Sciences, National Institutes of Health

Family-based designs enable assessment of genetic associations without bias from population stratification. However, parents are not always available - especially for diseases with onset later in life - and the case-sibling design, where each case is matched with one or more unaffected siblings, is useful. Analysis typically accounts for within-family dependencies by using conditional logistic

regression (CLR). We consider an alternative approach that treats each case-sibling set as a nuclear family with both parents missing by design. One can carry out maximum likelihood analysis by using the Expectation-Maximization (EM) algorithm to account for missing parental genotypes. We show that this approach improves power when some families have more than one unaffected sibling and also that under weak assumptions the approach enables the investigator to incorporate supplemental cases who do not have a sibling available and supplemental controls whose case sibling is not available (e.g. due to disability or death). We compare conditional logistic regression and the proposed missing parents approach under several risk scenarios. Our proposed method offers both improved statistical efficiency and asymptotically unbiased estimation for genotype relative risks and genotype-by-exposure interaction parameters.

e-mail: shi2@niehs.nih.gov

TWO-PHASE STUDIES OF GENE-ENVIRONMENT INTERACTION

Bhramar Mukherjee*, University of Michigan
Jaeil Ahn, University of Texas MD Anderson Cancer Center

Two-phase studies have been used as a cost and resource-saving alternative to classical cohort studies. For prioritizing individuals in an existing cohort for collecting additional genotype or environmental data this design seems particularly appealing for studies of gene-environment interaction. We will present a case-study on colorectal cancer where Phase I is a large case-control study base and exposure enriched sampling was implemented to select individuals for genotyping in Phase II. We will study various design and analysis choices in this general framework with a special focus on adaptive use of the gene-environment independence assumption at the design and inference stage. Moreover, we consider not just interaction parameters but the effect of this design for characterizing the joint or subgroup effects of the genetic or environmental factor.

e-mail: bhramar@umich.edu

METHODS FOR ANALYZING MULTIVARIATE PHENOTYPES IN GENE-BASED ASSOCIATION STUDIES USING FAMILIES

Saonli Basu*, University of Minnesota
Yiwei Zhang, University of Minnesota
Matt McGue, University of Minnesota

Gene-based genome-wide association studies (GWAS) provide a powerful alternative to the traditional single SNP association studies due to its substantial reduction in the multiple testing burden as well as possible gain in power due to modeling multiple SNPs within a gene. Implementing such gene-based association at a genome-wide level to detect association with multivariate traits present substantial analytical and computational challenges, especially for family-based designs. Recently, the canonical correlation analysis (CCA) has been gaining popularity due to its flexibility to test for association between multiple SNPs and multivariate traits. We have utilized the equivalence between the CCA and multivariate multiple linear regression (MMLR) to propose a rapid implementation of MMLR approach (RMMLR) in families. We compare through extensive simulation several gene-based association analysis approaches for both single and multivariate traits. Our RMMLR maintains valid type-I error even for genes with SNPs in strong LD. It also has substantial power for detecting genes in partial association with the correlated traits. We have also studied their performance on Minnesota Center for Twin Family Research dataset. In summary, our proposed RMMLR approach is an efficient and powerful technique to perform gene-based GWAS with single or multiple correlated traits.

e-mail: saonli@umn.edu

98. LARGE DATA VISUALIZATION AND EXPLORATION

VISUALIZING BRAIN IMAGING IN INTERACTIVE 3D

John Muschelli*, Johns Hopkins Bloomberg School of Public Health
Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Current research in neuromaging commonly presents results as orthogonal slices or as 3D-rendered static objects as journal article figures. These figures inherently discard the nature of the data, as well as eliminate the ability to view 4D data, whether it be through time or displaying multiple brain structures. We propose methods that allow researchers to easily present brain maps with the ability to interactively control the image by embedding them in a pdf document or publishing an interactive website. 3D Slicer is a useful tool for rendering 3D brain and region of interest (ROI) data that can be rendered online. This allows the user to interact with 3D/4D manipulatable objects without any hurdles of programming or third-party downloads. We show how in a matter of minutes, a researcher can have end-analysis figures to either share with collaborators or used as a journal figure. We also show an option to embed these 3D figures directly into a journal-ready pdf using misc3d in R.

e-mail: jmuschel@jhsph.edu

BIG DATA VISUALISATION IN R WITH GGPLOT2

Hadley Wickham*, RStudio

In this talk I will discuss recent work with extending ggplot2 to deal with very large datasets. The talk will discuss both the computational challenges of dealing with 100,000,000s of records, and the visualisation challenges of displaying that much data effectively.

e-mail: h.wickham@gmail.com

NETWORK VISUALISATION FOR PLAYFUL BIG DATA ANALYSIS

Amy R. Heineike*, Quid Inc

Paralleling its use in bioinformatics, large scale network visualisation is also a powerful tool for exploring unstructured human language corpora including of media streams and official government filings. Making data accessible, interactive and visually compelling allows us to create a playful paradigm of learning and exploration that allows users to make use of massive datasets in strategic analysis and decision making.

e-mail: amy.heineike@gmail.com

INTERACTIVE GRAPHICS FOR HIGH-DIMENSIONAL GENETIC DATA

Karl W. Broman*, University of Wisconsin, Madison

The value of interactive, dynamic graphics for making sense of high-dimensional data has long been appreciated but is still not in routine use. I will describe my efforts to develop interactive graphical tools for applications in genetics, using JavaScript and D3. I will focus on an expression genetics experiment in the mouse, with gene expression microarray data on each of six tissues, plus high-density genotype data, in each of 500 mice. I argue that in research with such data, precise statistical inference is not so important as data visualization.

e-mail: kbroman@biostat.wisc.edu

99. STATISTICAL ANALYSIS OF BIOMARKER INFORMATION IN NUTRITIONAL EPIDEMIOLOGY

CHALLENGES IN THE ANALYSIS OF BIOMARKER DATA FOR NUTRITION EPIDEMIOLOGY

Alicia L. Carriquiry*, Iowa State University

The incidence and severity of many chronic diseases can be linked to nutritional status. Until now, the nutritional status of population sub-groups has been evaluated using information from consumption surveys. By comparing nutrient intake with nutrient requirement, researchers and policy makers estimate the prevalence of inadequate or excessive intake and make recommendations. It is recognized, however, that for some nutrients, status is affected by factors other than intake. For example, vitamin D status is critically influenced by exposure to sun light. For others (e.g., sodium) it is difficult o accurately measure intake. The use of biomarker information is promising, but introduces its own challenges for analysis and interpretation of results. In this talk, we distinguish between biomarkers of status and biomarkers of diseases and discuss methods for analyzing information of

biomarkers for status. The remaining talks in the session address some of the methodological challenges that arise when analyzing and interpreting this type of information in the context of nutrition epidemiology.

e-mail: alicia@iastate.edu

BIOMARKERS OF NUTRITIONAL STATUS: METHODOLOGICAL CHALLENGES

Victor Kipnis*, National Cancer Institute, National Institutes of Health

Studies in nutritional epidemiology often fail to produce consistent associations between dietary intake and chronic disease. One of the main reasons may be substantial measurement error, both random and systematic, in self-reported assessment of dietary consumption. In the absence of true observed dietary intakes, statistical methods for adjusting for this error usually require a substudy with additional unbiased measurements of intake. In the talk, I will consider three classes of objective biomarkers of dietary consumption and discuss challenges involved in their use for mitigating the effect of dietary measurement error.

e-mail: kipnisv@mail.nih.gov

A SEMIPARAMETRIC APPROACH TO ESTIMATION IN MEASUREMENT ERROR MODELS WITH ERROR-IN-THE-EQUATION: APPLICATION TO SERUM VITAMIN D

Maria L. Joseph*, Iowa State University
Alicia L. Carriquiry, Iowa State University
Wayne A. Fuller, Iowa State University
Christopher T. Sempos, Office of Dietary Supplements, National Institutes of Health
Bess Dawson-Hughes, Human Nutrition Research Center on Aging at Tufts University

The nonlinear relationship between observed serum intact parathyroid hormone (iPTH) and observed serum 25-hydroxyvitamin D (25(OH)D) has been studied comprehensively. Many studies ignore the measurement error in these observed quantities. We use a nonlinear function to model the relationship between usual iPTH and usual 25(OH)D, where usual represents the long run average of the daily observations of these quantities. A semiparametric maximum likelihood approach is proposed to estimate the nonlinear relationship in a measurement error model with error-in-the-equation. The estimation procedures are applied to sample data.

e-mail: emme.jay11@gmail.com

IMPLEMENTATION OF A BIVARIATE DECONVOLUTION APPROACH TO ESTIMATE THE JOINT DISTRIBUTION OF TWO NON-NORMAL RANDOM VARIABLES OBSERVED WITH MEASUREMENT ERROR

Alicia L. Carriquiry, Iowa State University
Guillermo Basulto-Elías, Iowa State University
Eduardo A. Trujillo-Rivera*, Iowa State University

Replicate observations of 25(OH)D (biomarker for vitamin D status) and iPTH are available on a sample of individuals. We assume that measurements are subject to non-normal measurement error. We estimate the joint density of these bivariate data via non-parametric deconvolution. The estimated density is used to compute statistics of public health interest, such as the proportion of persons in a group with 25(OH)D values below iPTH, or the value of 25(OH)D above which iPTH is approximately constant. We use a bootstrap approach to compute confidence intervals. Several bivariate kernel density estimators for the noisy data and estimators for the characteristic function of the error are compared.

e-mail: eduardo@iastate.edu

100. UTILITIES OF STATISTICAL MODELING AND SIMULATION FOR DRUG DEVELOPMENT

GUIDED CLINICAL TRIAL DESIGN: DOES IT IMPROVE THE FINAL DESIGN?

J. Kyle Wathen*, Janssen Research & Development

Many important details of a clinical trial are often ignored in order to simplify the statistical design. In this talk I will present a case study of a trial where simulation was used to gain insight into the impact of various adaptations such as 2-stage design vs a single stage design, safety rules based on two correlated outcomes and futility/superiority decisions based on a third outcome. Simulation was also used to investigate the impact of many design considerations as well as statistical modeling. Through the use of simulation several logistical and statistical issues were raised, however, did the simulation improve the final clinical trial design?

e-mail: kwathen@its.jnj.com

ON THE CHOICE OF DOSES FOR PHASE III CLINICAL TRIALS

Carl-Fredrik Burman*, AstraZeneca Research & Development

It is important to consider how to optimize the choice of dose or doses that continue into the confirmatory phase. Phase IIB dose-finding trials are relatively small and often lack the ability of precisely estimating the dose-response curves for efficacy and tolerability. Using simple but illustrative models, we find the optimal doses and compare the probability of success, for fixed total sample sizes, when one or two active doses are included in phase III.

e-mail: carl-fredrik.burman@astrazeneca.com

SIMULATION-GUIDED DESIGN FOR MOLECULARLY TARGETED THERAPIES IN ONCOLOGY

Cyrus R. Mehta*, Cytel Inc.

The development of molecularly targeted therapies for certain types of cancers (e.g., Vemurafenib for advanced melanoma with mutant BRAF; Cetuximab for metastatic colorectal cancer with KRAS wild type) has let to the consideration of population enrichement designs that explicitly factor-in the possibility that the experimental compound might differentially benefit different biomarker subgroups. In such designs, enrollment would initially be open to a broad patient population with the option to restrict future enrollment, following an interim analysis, to only those biomarker subgroups that appeared to be benefiting from the experimental therapy. While this strategy could greatly improve the chances of success for the trial, it poses several statistical and logistical design challenges. Since late-stage oncology trials are typically event driven, one faces a complex trade-off between power, sample size, number of events and study duration. This trade-off is further compounded by the importance of maintaining statistical independence of the data before and after the interim analysis and of optimizing the timing of the interim analysis. This talk will highlight the crucial role of simulation-guided design for resolving these difficulties while nevertheless maintaining strong control of the type-1 error.

e-mail: mehta@cytel.com

101. RECENT ADVANCES IN SURVIVAL AND EVENT-HISTORY ANALYSIS

ANALYSIS OF DIRECT AND INDIRECT EFFECTS IN SURVIVAL ANALYSIS

Odd O. Aalen*, University of Oslo, Norway

Mediation analysis of survival data has been sorely missing. In many settings one runs Cox analyses with baseline covariates while internal time-dependent covariates are not included in the analysis due to perceived difficulties of interpreting results. At the same time it is clear that the time-dependent covariates may contain information about the mechanism of the treatment effects. We shall discuss the use of dynamic path analysis for studying this issue. The relation to causal inference will be pointed out.

e-mail: o.o.aalen@medisin.uio.no

RECURRENT MARKER PROCESSES WITH COMPETING TERMINAL EVENTS

Mei-Cheng Wang*, Johns Hopkins Bloomberg School of Public Health

In follow-up or surveillance studies, marker measurements are frequently collected or observed conditioning on the occurrence of recurrent events. In many situations, the marker measurement exists only when a recurrent event took place. Examples include medical cost for inpatient or outpatient cares, length-of-stay for hospitalizations, and prognostic measurements repeatedly measured at incidences of infection. A recurrent marker process, defined between an initiating event and a terminal event, is composed of recurrent events and markers. This talk considers the situation when the occurrence of terminal event is subject to competing risks. Statistical methods and inference of recurrent marker process are developed to address a variety of questions/applications for the purposes of estimating and comparing (i) real-life utility measures, such as medical cost or length-of-stay in hospital, for different competing risk groups, and (ii) recurrent marker processes for different treatment plans in relation to competing risk groups. A SEER-Medicare-linked data base is used to illustrate the proposed approaches.

e-mail: mcwang@jhsph.edu

DISTRIBUTION-FREE INFERENCE METHODS FOR THRESHOLD REGRESSION

G. A. (Alex) Whitmore*, McGill University
Mei-Ling T. Lee, University of Maryland

This research considers a survival model in which failure (or other endpoint) is triggered when a stochastic process first reaches a critical threshold or boundary. Parameters of the threshold, process, and time scale are estimated from censored survival data, possibly augmented by process levels for survivors. Co-variates are handled using threshold regression as described in Lee and Whitmore (Statistical Sciences, 2006, 501-513). Individual outcomes in this setting are observation pairs (U, V) where U is a change in process level and V is a change in time. The outcomes are of two kinds: (1) If the individual survives until the end of the study, U is random and V is fixed; (2) If the individual fails during the study, U is fixed (except for threshold overshoot) and V is random. We develop a Wald-type identity for (U, V) for a wide-class of stochastic processes that allows a distribution-free approach to statistical inference. We present the mathematical theory and a suite of estimation and prediction methods and illustrate them with medical case applications.

e-mail: george.whitmore@mcgill.ca

ESTIMATING THE COUNTING STATISTICS OF A SELF-EXCITING PROCESS

Paula R. Bouzas*, University of Granada, Spain
Nuria Ruiz-Fuentes, University of Jaén, Spain

Recurrent event data are event history data when the information of the occurrences times is available. It is common to model this type of data by counting processes. A self-exciting process is a counting process with memory because its intensity process depends on the past of the counting one. The count-conditional intensity characterizes the counting process since the probability mass function and some other statistics are expressed in terms of it. This work presents the estimation of the count-conditional intensity when the data are observed as recurrent event data. Afterwards, it is used to estimate counting statistics such as the probability mass function and the mean of the self-exciting process. Simulations illustrate these estimations.

e-mail: paula@ugr.es

102. INNOVATIVE METHODS IN CAUSAL INFERENCE WITH APPLICATIONS TO MEDIATION, NEUROIMAGING, AND INFECTIOUS DISEASES

BAYESIAN CAUSAL INFERENCE FOR MULTIPLE MEDIATORS

Michael Daniels*, University of Texas, Austin
Chanmin Kim, University of Florida

In behavioral studies, the causal effect of a intervention is of interest to researchers. There have been many approaches proposed for causal mediation analysis, but mostly for the single mediator case. This is due in part to causal interpretations of multiple mediators being quite complex both in terms of identifying and interpreting appropriate causal effects. Most of these approaches rely on a sequential ignorability and no-interaction assumptions, which can be hard to justify in behavioral trials. Here, we propose a Bayesian approach to infer natural direct and indirect effects of multiple mediators. Our approach avoids the sequential ignorability assumption and allows for estimation of the indirect effects of individual mediators and the joint effects of multiple mediators.

e-mail: mjdaniels@austin.utexas.edu

INFERENCE WITH INTERFERENCE IN fMRI

Xi Luo*, Brown University
Dylan S. Small, University of Pennsylvania
Chiang-shan R. Li, Yale University
Paul R. Rosenbaum, University of Pennsylvania

An experimental unit is an opportunity to randomly apply or withhold a treatment. There is interference between units if the application of the treatment to one unit may also affect other units. In cognitive neuroscience, a common form of experiment presents a sequence of stimuli or requests for cognitive activity at random to each experimental subject and measures biological aspects of brain activity that follow these requests. Each subject is then many experimental units, and interference between units within an experimental subject is, likely, in part because the stimuli follow one another quickly and in part because human subjects learn or become experienced or primed or bored as the experiment proceeds. In this talk, we describe and further develop methodology for inferring treatment effects in the presence of interference. This method employs nonparametric placement tests. Its effectiveness is illustrated using a functional magnetic resonance imaging (fMRI) experiment concerned with the inhibition of motor activity. A simulation study evaluates the power of competing procedures.

e-mail: xi.rossi.luo@gmail.com

ASSESSING THE EFFECTS OF CHOLERA VACCINATION IN THE PRESENCE OF INTERFERENCE

Michael G. Hudgens*, University of North Carolina, Chapel Hill

Interference occurs when the treatment of one person may affect the outcome of another. For example, in infectious diseases, whether one individual is vaccinated may affect whether another individual becomes infected or develops disease. Quantifying such indirect (or spillover) effects of vaccination can have important public health or policy implications. In this talk we apply recently developed inverse-probability weighted (IPW) estimators of treatment effects in the presence of interference to an individually-randomized, placebo controlled trial of cholera vaccination in 121,982 individuals in Matlab, Bangladesh. Because these IPW estimators have not been employed previously, a simulation study was also conducted to assess the empirical behavior of the estimators in settings similar to the cholera vaccine trial. Simulation study results demonstrate the IPW estimators can yield unbiased estimates of the direct, indirect, total and overall effects of treatment in the presence of interference. Application of the IPW estimators to the

cholera vaccine trial indicates a significant indirect effect of vaccination. For example, among placebo recipients the incidence of cholera was reduced by 5.5 cases per 1000 individuals, (95% CI: 3.2, 7.7) in neighborhoods with 60% vaccine coverage compared to neighborhoods with 33% coverage.

e-mail: mhudgens@bios.unc.edu

103. CLINICAL TRIALS

A MULTISTAGE NON-INFERIORITY STUDY ANALYSIS PLAN TO EVALUATE SUCCESSIVELY MORE STRINGENT CRITERIA FOR A CLINICAL TRIAL WITH RARE EVENTS

Siying Li*, University of North Carolina, Chapel Hill
Gary G. Koch, University of North Carolina, Chapel Hill

We address a multistage clinical trial to assess a sequence of hypotheses in the non-inferiority and also rare events setting. Three successive hypotheses are used to evaluate whether the new treatment meets the criteria for new drug approval. Sample sizes for a five stage trial for all hypotheses are calculated using Poisson and Logrank sample size methods. Three strategies and corresponding analysis plans are developed to evaluate the sequential hypotheses. Simulations show the design is satisfactory with respect to controlled Type I error, sufficient power, and early success at interim analyses.

e-mail: siying@live.unc.edu

ON THE EFFICIENCY OF NONPARAMETRIC VARIANCE ESTIMATION IN SEQUENTIAL DOSE-FINDING

Chih-Chi Hu*, Columbia University Mailman School of Public Health
Ying Kuen K. Cheung, Columbia University Mailman School of Public Health

Typically, phase I trials are designed to determine the maximum tolerated dose, defined as the maximum test dose that causes a toxicity with a target probability. In this talk, we formulate dose finding as a quantile estimation problem and focus on situations where toxicity is defined by dichotomizing a continuous outcome, for which a correct specification of the variance function of the outcomes is important. This is especially true for sequential study where the variance assumption directly involves in the generation of the design points and hence sensitivity analysis may not be performed after the data are collected. In this light, there is a strong reason for avoiding parametric assumptions on the variance function, although this may incur efficiency loss. We investigate how much information one may retrieve by making additional parametric assumptions on the variance in the context of a sequential least squares recursion. By asymptotic comparison and simulation study, we demonstrate that assuming homoscedasticity achieves only a modest efficiency gain when compared to nonparametric variance estimation: when homoscedasticity in truth holds, the latter is at worst 88% as efficient as the former in the limiting case, and often achieves well over 90% efficiency for most practical situations.

e-mail: fsnycfan@gmail.com

BAYESIAN ENROLLMENT AND STOPPING RULES FOR MANAGING TOXICITY REQUIRING LONG FOLLOW-UP IN PHASE II ONCOLOGY TRIALS

Guochen Song*, Quintiles
Anastasia Ivanova, University of North Carolina, Chapel Hill

Stopping rules for toxicity are routinely used in phase II oncology trials. If the follow-up for toxicity is long, it is desirable to have a stopping rule that uses all toxicity information available not only information from patients with full follow-up. Further, to prevent excessive toxicity in such trials we propose an enrollment rule that informs an investigator about the maximum number of patients that can be enrolled depending on current enrollment and all available information about toxicity. We give recommendations on how to construct Bayesian stopping and enrollment rules to monitor toxicity continuously in Phase II oncology trials with a long follow-up.

e-mail: guochens@gmail.com

ANALYSIS OF SAFETY DATA IN CLINICAL TRIALS USING A RECURRENT EVENT APPROACH

Qi Gong*, Amgen Inc.
Yansheng Tong, Genentech Inc.
Alexander Strasak, F. Hoffmann-La Roche Ltd.
Liang Fang, Genentech Inc.

As an important aspect of the clinical evaluation of an investigational therapy, safety data are routinely collected in clinical trials. To date, the analysis of safety data has largely been limited to descriptive summaries of incidence rates, or contingency tables aiming to compare simple rates between treatment arms. Many have argued this traditional approach failed to take into account important information including severity, onset time, and duration of a safety signal. In this article, we propose a framework to summarize safety data with mean frequency function and compare safety profiles between treatments with a generalized log-rank test, taking into account the aforementioned characteristics ignored in traditional analysis approaches. In addition, a multivariate generalized log-rank test to compare the overall safety profile of different treatments is proposed. In the proposed method, safety events are considered to follow a recurrent event process with a terminal event for each patient. The terminal event is modeled by a process of two types of competing risks: safety events of interest and other terminal events. Statistical properties of the proposed method are investigated via simulations. An application is presented with data from a phase II oncology trial.

e-mail: gongqi@gmail.com

SMALLER, FASTER PHASE III TRIALS: A BETTER WAY TO ASSESS TARGETED AGENTS?

Karla V. Ballman*, Mayo Clinic
Marie-Cecile Le Deley, Institut Gustave Roussy, Université Paris-Sud 11
Daniel J. Sargent, Mayo Clinic

Traditional clinical trial designs aim to definitively establish the superiority, which results in large sample sizes. Increasingly, common cancers are recognized to consist of small subgroups making large trials infeasible. We compared trial design strategies with different combinations of sample size and alpha to determine which performs best over a 15 yr research horizon. We simulated a series of two-treatment superiority trials using different values for the alpha-level and trial sample size (SS). Different disease scenarios, accrual rates, and distributions of treatment effects were used. Metrics used included: impact on hazard ratio (comparing yr 15 vs yr 0), overall survival benefit (difference in median survival between year 15 and year 0), and risk of worse survival at year 15 compared to year 0. Overall survival gains were greater as alpha increased from 0.025 to 0.20. Gains in survival were achieved with SSs smaller than required under traditional criteria. Reducing the SS and increasing alpha increased the likelihood of having a poorer survival rate at yr 15, but this probability remained small. Results were consistent under different assumed distributions for treatment effect. As patient populations become more restricted (and thus smaller), the current risk adverse trial design strategy may slow long term progress and deserves re-examination.

e-mail: ballman@mayo.edu

SUPERIORITY TESTING IN GROUP SEQUENTIAL NON-INFERIORITY TRIALS

Vandana Mukhi*, U.S. Food and Drug Administration
Heng Li, U.S. Food and Drug Administration

In non-inferiority clinical trials it is often of interest to pre-specify that if the non-inferiority null hypothesis is rejected then superiority will be tested. In group-sequential non-inferiority trials, this pre-specification would need to contain not only a decision boundary associated with the non-inferiority hypothesis, but also a rejection rule for the superiority null hypothesis at interim and final stages. We will consider some design issues in this setup, in particular type I error rate for the superiority test.

e-mail: vandana26@yahoo.com

COMPARING STUDY RESULTS FROM VARIOUS PROPENSITY SCORE METHODS USING REAL CLINICAL TRIAL DATA

Terri K. Johnson*, U.S. Food and Drug Administration
Yunling Xu, U.S. Food and Drug Administration

Often a randomized trial is deemed unfeasible or unethical for evaluating the safety and effectiveness of some medical devices. In such cases, a non-randomized trial that utilizes a historical control data may be conducted when sufficient clinical knowledge and information are available. Propensity score methods are often used to reduce biases due to unbalanced covariates. Although many literatures have discussed three common techniques which utilize the propensity scores (matching, stratification, and regression methods) to evaluate the treatment effect, how results from these three methods vary in a regulatory setting has not been well studied. We will examine and describe the use of propensity score methods in medical device studies, and compare the study results by applying various propensity score methods to PMA data.

e-mail: terri.johnson@fda.hhs.gov

104. NEXT GENERATION SEQUENCING

DELPTM: A STATISTICAL ALGORITHM TO IDENTIFY POST-TRANSLATIONAL MODIFICATIONS FROM TANDEM MASS SPECTROMETRY (MS/MS) DATA

Susmita Datta*, University of Louisville
Jasmit S. Shah, University of Louisville

Post translational modification (PTM) of a protein plays a significant role in complex diseases such as cancer and diabetes. Hence, the identification of PTM on a genome-wide scale is important. It is possible to identify PTMs through analysis of tandem mass spectrometry data of disease affected fluids or tissues. Identification of PTM with unrestrictive blind search algorithm is most helpful at the exploratory stage of a research when it is unknown to restrict the search within a known class of PTMs. However, these methods suffer from mass measurement inaccuracy and uncertainty in predicting modification positions. Here in this work we propose to modify the results of any blind search algorithm through statistical methodology. We develop a self-validated clustering method for mixed data types through rank aggregation of the PTM data. We then use the number of clusters as known groups of the PTM data and subsequently use Bayesian modeling to define the likelihood of the data with the knowledge of the chemical process of PTM.

e-mail: susmita.datta@louisville.edu

PROTEIN IDENTIFICATION: A BAYESIAN APPROACH

Nicole Lewis*, University of South Carolina
David B. Hitchcock, University of South Carolina
Ian L. Dryden, University of Nottingham
John R. Rose, University of South Carolina

Current methods for protein identification in tandem mass spectrometry involve database searches or de novo peptide sequencing. With database searches, there is a limitation involving the method due to the relatively low number of known proteins. Shortcomings of de novo peptide sequencing include incomplete b and y ion sequences and lack of accuracy of the formed peptides. Here we present a Bayesian approach to identifying peptides. Our model uses prior information about the average relative abundances of bond cleavages and the prior probability of any particular amino acid sequence. The proposed likelihood function is comprised of two overall distance measures, which measure how close an observed spectrum is to a theoretical scan for a peptide. A Markov chain Monte Carlo algorithm is employed to simulate candidate choices from the posterior distribution of the peptide. The true peptide is estimated as the peptide with the largest posterior density. In addition, our method is designed to rank top candidate peptides according to their approximate posterior densities, which allows one to see the relative uncertainty in the best choice. Our method is not dependent upon known peptides as in the database searches and aims to alleviate some of the drawbacks of de novo sequencing.

e-mail: nicole_lewis8@hotmail.com

iASeq: INTEGRATIVE ANALYSIS OF ALLELE-SPECIFICITY OF PROTEIN-DNA INTERACTIONS IN MULTIPLE ChIP-seq DATASETS

Yingying Wei*, Johns Hopkins University Bloomberg School of Public Health
Xia Li, Chinese Academy of Sciences
Qianfei Wang, Chinese Academy of Sciences
Hongkai Ji, Johns Hopkins University Bloomberg School of Public Health

ChIP-seq provides new opportunities to study allele-specific protein-DNA binding (ASB). Currently, little is known about the correlation patterns of allele-specificity among different transcription factors and epigenetic marks. Moreover, detecting allelic imbalance from a single ChIP-seq dataset often has low statistical power since only sequence reads mapped to heterozygote SNPs are informative for discriminating two alleles. We develop a new method iASeq to address both issues by jointly analyzing multiple ChIP-seq datasets. iASeq uses a Bayesian hierarchical mixture model to identify correlation patterns of allele-specificity among multiple proteins. Using the discovered correlation patterns, the model allows one to borrow information across datasets to improve detection of allelic imbalance. Application of iASeq to 77 ChIP-seq samples from 40 ENCODE datasets and 1 Genomic DNA sample in GM12878 cells reveals that allele-specificity of multiple proteins are highly correlated. The analysis also demonstrates the ability of iASeq to improve allelic inference compared to analyzing each individual dataset separately. iASeq illustrates the value of integrating multiple datasets in the allele-specificity inference and offers a new tool to better analyze ASB.

e-mail: ywei@jhsph.edu

DIFFERENTIAL EXPRESSION ANALYSIS OF RNA-seq DATA AT BASE-PAIR RESOLUTION

Alyssa C. Frazee*, Johns Hopkins University
Rafael Irizarry, Johns Hopkins University
Jeffrey T. Leek, Johns Hopkins University

As the cost of generating and storing RNA sequencing data has decreased, this high-throughput gene expression data has become much more abundant. Because this type of data is very useful in analyzing differential expression, it has become clear that demand is high for a differential expression analysis pipeline that (a) does not depend on existing annotation and (b) does not require transcriptome assembly. We have developed a pipeline that fits these criteria. The proposed method first tests for differential expression at each base-pair in the genome, and then groups consecutive base-pairs with the same differential expression classification into regions. These results can be used to make inference about differentially expressed exons and genes, or they can enable identification of previously un-annotated transcribed regions. This method has been implemented in a user-friendly, open-access software package for analysis and visualization of the results.

e-mail: afrazee@jhsph.edu

A NOVEL FUNCTIONAL PCA METHOD FOR TESTING DIFFERENTIAL EXPRESSION WITH RNA-seq DATA

Hao Xiong*, University of California, Berkeley
Haiyan Huang, University of California, Berkeley
Peter Bickel, University of California, Berkeley

RNA-Seq provides multi-resolution views of transcriptome complexity: at exon, SNP, and positional level; splicing; post-transcriptional RNA editing; isoform and allele-specific expression. But despite the unsurpassed resolution, current statistical methods for testing differential expressions only compare single-number summaries of genes across experimental conditions. This leaves RNA-Seq analysis vulnerable to sequencing errors, nucleotide composition effects, alternative splicing, allele specific expressions, and composition of raw sequence. There is no consensus on a standard and comprehensive approach to correct biases and reduce noise. We propose a two-parameter generalized nonhomogeneous Poisson model to characterize base-level counts, whose parameters vary with bases. We incorporate base-specific

variation into the model. The new model performs more reasonable normalization and offer more accurate estimation of base-level expression. The corrected expression can be modelled as random functions and be expanded in orthogonal functional principal components through Karhunen-Loeve decomposition. Testing differential expressions here is comparing FPCA scores, instead of difference in read-counts of gene. The proposed methods are applied to drosophila and schizophrenia and bipolar RNA-Seq datasets.

e-mail: xiongha@gmail.com

CAN HUMAN ETHNIC SUBGROUPS BE UNCOVERED BY NEXT GENERATION SEQUENCING DATA?

Yiwei Zhang*, University of Minnesota
Wei Pan, University of Minnesota

Population stratification is of primary interest in genetic studies to imply human evolution history and to avoid spurious findings in association testing. Next generation sequencing data brings greater chance as well as challenges to uncover population structure in finer scales. For SNP data, the most commonly used method is principal component analysis (PCA), while two recently proposed methods are Spectral Clustering (Spectral-GEM) and Locally Linear Embedding (LLE). In this talk we apply and compare these three methods using the whole-genome sequence data of the European and African samples from the 1000 Genomes Project to uncover ethnic subgroups.

e-mail: zhan1447@umn.edu

AUTOREGRESSIVE MODELING AND VARIABLE SELECTION PROCEDURES IN HIDDEN MARKOV MODELS WITH COVARIATES, WITH APPLICATIONS TO DAE-seq DATA

Naim U. Rashid*, University of North Carolina, Chapel Hill
Wei Sun, University of North Carolina, Chapel Hill
Joseph G. Ibrahim, University of North Carolina, Chapel Hill

In DAE (DNA After Enrichment)-seq experiments, DNA related with certain biological processes are isolated and sequenced on a high-throughput sequencing platform to determine their genomic positions. Statistical analysis of DAE-seq data aims to detect genomic regions with significant aggregations of isolated DNA. However, several confounding factors and their interactions may bias DAE-seq data, which leads to a challenging variable selection problem. In addition, signals in adjacent genome regions may exhibit strong correlations, invalidating the independence assumption of many existing methods for DAE-seq data analysis. To mitigate these issues, we develop a novel Autoregressive Hidden Markov Model (AR-HMM) accounting for covariate effects and violations of the independence assumption. We demonstrate that our AR-HMM leads to improved performance in identifying enriched regions in both simulated and real datasets, especially in those with broader regions of DAE-seq signal enrichment. We also introduce a variable selection procedure in the context of the HMM when the mean of each state-specific emission distribution is modeled by some set of covariates. We study the theoretical properties of this variable selection method and demonstrate its efficacy in simulated and real DAE-seq data.

e-mail: naim@unc.edu

105. NONPARAMETRIC METHODS

VARIABLE SELECTION IN MONOTONE SINGLE-INDEX MODELS VIA THE ADAPTIVE LASSO

Jared Foster*, University of Michigan

We consider the problem of variable selection for monotone single-index models. A single-index model assumes that the expectation of the outcome is an unknown function of a linear combination of covariates. Assuming monotonicity of the unknown function is often reasonable, and allows for more straightforward inference. We present an adaptive LASSO penalized least squares approach to estimating the index parameter and the unknown function in these models for continuous outcome. Monotone function estimates are achieved using the pooled adjacent violators algorithm, followed by kernel regression. In the iterative estimation process, a linear approximation to the unknown function is used, therefore reducing the situation to that of linear regression, and allowing for the use of standard LASSO algorithms, such as coordinate descent. Results of a simulation study indicate that the proposed methods perform well under a variety of circumstances, and that an assumption of monotonicity, when appropriate, noticeably improves performance. The proposed methods are applied to data from a randomized clinical trial for the treatment of a critical illness in the intensive care unit (ICU).

e-mail: jaredcf@umich.edu

CROSS-VALIDATION AND A U-STATISTIC MODEL SELECTION TOOL

Qing Wang*, Williams College
Bruce G. Lindsay, The Pennsylvania State University

In this talk we turn our attention to the problem of model selection and will propose an alternative model selection method, akin to the BIC criterion. We construct a U-statistic form estimate for the likelihood risk that is the basis of the generalized AIC methods. The U-statistic risk estimate, sometimes called likelihood cross-validation, is an alternative estimator to the generalized AIC and is equivalent to the BIC method when the subsample size equals to n/(log n-1). The proposed cross-validation methodology is more generally applicable than AIC and BIC. In addition, with an appropriate estimate for the variance of a general U-statistic, one can test which model has the smallest risk based on the proposed U model selection tool. A real data example is provided to study our estimator. In addition to determining the lowest risk model in the BIC sense, we compare the proposed U-statistic cross-validation tool with the standard criteria.

e-mail: qing.w.wang@williams.edu

MAXIMUM LIKELIHOOD ESTIMATION FOR SEMIPARAMETRIC EXPONENTIAL TILT MODELS WITH ADJUSTMENT OF COVARIATES

Jinsong Chen*, University of Illinois at Chicago
George R. Terrell, Virginia Tech University
Inyoung Kim, Virginia Tech University

We propose a semiparametric exponential tilt model allowing the adjustment of covariates. Furthermore, we add flexible log-concave qualitative constraint on nonparametric density estimation of proposed model. Maximum likelihood method is used for estimate exponential tilt parameters and density functions. Asymptotic normality of the estimates is developed. Likelihood ratio test, which is proved to follow a chi-square distribution, is constructed to test the significance of exponential tilt parameter estimation. Simulation study is conducted to assess the performance of our method. Our model is also applied to analyze the data from Chicago Healthy Aging Study.

e-mail: jschen24@hotmail.com

TWO STEP ESTIMATION OF PROPORTIONAL HAZARDS REGRESSION MODELS WITH NONPARAMETRIC ADDITIVE EFFECTS

Rong Liu*, University of Toledo

The Cox proportional hazards model usually assumes that the covariate has a log-linear effect on the hazard function. Many studies had been done for removing the linear restriction. Sleeper and Harrington (1990) used additive models to model the nonlinear covariate effects in the Cox model. But the asymptotic properties of the estimation of component functions were not obtained. We propose spline-backfitted kernel (SBK) estimator for the component functions, and establish oracle properties for the two-step estimator of each function component such that it performs as well as the univariate function estimator by assuming that all other function components are known. Asymptotic distributions and consistency properties of the estimators are obtained. Simulation evidence strongly corroborates with the asymptotic theory. We illustrate the method with a real data example.

e-mail: rong.liu@utoledo.edu

TWO-SAMPLE DENSITY-BASED EMPIRICAL LIKELIHOOD RATIO TESTS BASED ON PAIRED DATA, WITH APPLICATION TO A TREATMENT STUDY OF ATTENTION-DEFICIT/HYPERACTIVITY DISORDER AND SEVERE MOOD DYSREGULATION

Albert Vexler*, The State University of New York at Buffalo

It is a common practice to conduct medical trials in order to compare a new therapy with a standard-of-care based on paired data consisted of pre- and post-treatment measurements. In such cases, a great interest often lies in identifying treatment effects within each therapy group as well as detecting a between-group difference. In this article, we propose exact nonparametric tests for composite hypotheses related to treatment effects to provide efficient tools that compare study groups utilizing paired data. When correctly specified, parametric likelihood ratios can be applied, in an optimal manner, to detect a difference in distributions of two samples based on paired data. The recent statistical literature introduces density-based empirical likelihood methods to derive efficient nonparametric tests that approximate most powerful Neyman-Pearson decision rules. We adapt and extend these methods to deal with various testing scenarios involved in the two-sample comparisons based on paired data. We show the proposed procedures outperform classical approaches. An extensive Monte Carlo study confirms that the proposed approach is powerful and can be easily applied to a variety of testing problems in practice. The proposed technique is applied for comparing two therapy strategies to treat children's attention deficit/hyperactivity disorder and severe mood dysregulation.

e-mail: avexler@buffalo.edu

ESTIMATING THE DISTRIBUTION FUNCTION USING RANKED SET SAMPLES FROM BIASED DISTRIBUTIONS

Kaushik Ghosh*, University of Nevada, Las Vegas
Ram C. Tiwari, U.S. Food and Drug Administration

In this work, we investigate the estimation of the underlying (unbiased) distribution using generalized ranked set samples from its various biased versions. We propose a nonparametric estimator of the distribution function and investigate its asymptotic properties. We also present results of simulation studies and illustrate our proposed method with a real data set.

e-mail: kaushik.ghosh@unlv.edu

A UNIFYING FRAMEWORK FOR RANK TESTS

Jan R. De Neve*, Ghent University
Olivier Thas, Ghent University and University of Wollongong
Jean-Pierre Ottoy, Ghent University

We demonstrate how well known rank tests, such as the Wilcoxon-Mann-Whitney, Kruskal-Wallis, and Friedman test, and many more, can be embedded in a statistical modeling methodology and how our approach can be used for constructing new rank tests for more complicated designs. In particular, rank tests for unbalanced and multi-factor designs, and rank tests that allow for correcting for continuous confounders are included. In addition to hypotheses testing, the method allows for the estimation of meaningful effect sizes, resulting in a better understanding of the data. Our method results from two particular parameterisations of Probabilistic Index Models (PIM).

e-mail: janr.deneve@ugent.be

106. JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

JOINT MODELING OF SURVIVAL DATA AND MISMEASURED LONGITUDINAL DATA USING THE PROPORTIONAL ODDS MODEL

Juan Xiong, University of Western Ontario
Wenqing He*, University of Western Ontario
Grace Yi, University of Waterloo

Joint modeling of longitudinal and survival data has been studied extensively, where the Cox proportional hazards model has frequently been used to incorporate the relationship between survival time and covariates. Although the proportional odds model is an attractive alternative to the Cox proportional hazards model by featuring the dependence of survival times on covariates via cumulative covariate effects, this model is rarely discussed in the joint modeling context. To fill this gap, we investigate joint modeling of the survival data and longitudinal data which subject to measurement error. We describe a model parameter estimation method based on expectation maximization algorithm. In addition, we assess the impact of naïve analyses that fail to address error occurring in longitudinal measurements. The performance of the proposed method is evaluated through simulation studies and a real data analysis.

e-mail: whe@stats.uwo.ca

JOINT MODELING OF LONGITUDINAL DATA AND INFORMATIVE OBSERVATIONAL TIMES WITH TIME-VARYING COEFFICIENTS

Liang Li*, Cleveland Clinic

Patients undergoing atrial fibrillation (AF) ablation surgery often receive repeated ECG exams in the post-operative period to determine if they are still at risk of AF. In an observational study, the time that the patient returns to the clinic to have ECG exams may be correlated with their risk of AF, and the association between baseline variables and risk of AF may differ between the early and late phases of post-operative recovery. To address these two challenges, we propose a generalized linear mixed model for a binary outcome variable, with time-varying regression coefficients; the observational times of the outcome variable are assumed to be correlated with the individual risk of AF through a random effect. The model parameters are estimated by maximizing the joint likelihood of the longitudinal binary outcome variable and the serial observational times, given the covariates. The time-varying coefficient functions are modeled by penalized splines, allowing for different smoothness for different coefficient functions. The random effect of the generalized linear mixed model and the spline coefficients form a two-stage hierarchical high dimensional random effect structure and are handled by numerical integration and Laplace approximation, respectively. Simulations and a real data application are presented to illustrate the empirical performance of the proposed method.

e-mail: lil2@ccf.org

A BAYESIAN APPROACH TO JOINT ANALYSIS OF PARAMETRIC ACCELERATED FAILURE TIME AND MULTIVARIATE LONGITUDINAL DATA

Sheng Luo*, University of Texas, Houston

Impairment caused by Parkinson's disease (PD) is multidimensional (e.g., sensoria, functions, and cognition) and progressive. Its multidimensional nature precludes a single outcome to measure disease progression. Clinical trials of PD use multiple categorical and continuous longitudinal outcomes to assess treatment effect on overall improvement. A terminal event such as death or dropout can stop the follow-up process. Moreover, the time to terminal event may be dependent on the multivariate longitudinal measurements. In this article, we consider a joint random-effects model for the correlated outcomes. A multilevel item response theory model is used for the multivariate longitudinal outcomes and a parametric accelerated failure time model is used for the failure time due to the violation of proportional hazard assumption. These two models are linked via random effects. The Bayesian inference via Markov Chain Monte Carlo is implemented in BUGS language. Our proposed method is evaluated by simulation studies and is applied to DATATOP study, a motivating clinical trial to determine if deprenyl slows the progression of PD.

e-mail: sheng.t.luo@uth.tmc.edu

PREDICTION ACCURACY OF LONGITUDINAL BIOMARKERS IN JOINT LATENT CLASS MODELS

Lan Kong*, The Pennsylvania State University College of Medicine
Guodong Liu, The Pennsylvania State University College of Medicine

Biomarkers are often measured over time in longitudinal studies and clinical trials to better understand the biological pathways of disease development and the mechanism of treatment effects. Joint modeling of longitudinal biomarker data and time-to-event data has been intensively studied, with the focus being the association between longitudinal process and primary endpoint. The use of joint models as prediction tools has gained increasing attention lately. The probability of experiencing an event of interest by a certain time point can be estimated for a future subject given the clinical information and available longitudinal biomarker measurements. Under the joint latent class modeling framework, we develop the prediction accuracy measures for evaluating the clinical utility of longitudinal biomarkers. In particular, we derive the discrimination and calibration measures for joint latent class models based on the recent work in this field. We further illustrate our method in predicting the subject-specific risk of progression to Alzheimer's disease with the magnetic resonance imaging and cerebrospinal fluid biomarker data from the Alzheimer's Disease Neuroimaging Initiative (ADNI).

e-mail: lkong@phs.psu.edu

STRUCTURAL NESTED MODELS FOR JOINT MODELING OF REPEATED MEASURES AND SURVIVAL OUTCOMES

Marshall M. Joffe*, University of Pennsylvania

We consider how to formulate structural nested models for the effect of a treatment sequence on outcomes consisting jointly consisting of a failure-time and a sequence of repeated measures when failure precludes observation of subsequent outcomes. We consider interpretation of the models and different approaches to estimation, including fully parametric, fully semiparametric, and a mixed approach, with a parametric approach for the failure-time outcome and a semiparametric approach to the repeated measures outcome.

e-mail: mjoffe@mail.med.upenn.edu

JOINT MODELING OF LONGITUDINAL AND CURE-SURVIVAL DATA

Sehee Kim*, University of Michigan
Donglin Zeng, University of North Carolina, Chapel Hill
Yi Li, University of Michigan
Donna Spiegelman, Harvard School of Public Health

This article presents semiparametric joint models to analyze longitudinal measurements and survival data with a cure fraction. We consider a broad class of transformations for the cure-survival model, which includes the popular proportional hazards cure models and the proportional odds cure models as special cases. We propose to estimate all the parameters using the nonparametric maximum likelihood estimators (NPMLE). We provide the simple and efficient EM algorithms via Laplace transformation to implement the proposed inference procedure. Asymptotic properties of the estimators are shown to be asymptotically normal and semiparametrically efficient. Finally, we demonstrate the good performance of the method through extensive simulation studies and a real-data application.

e-mail: seheek@umich.edu

REGRESSION MODELING OF LONGITUDINAL DATA WITH INFORMATIVE OBSERVATION TIMES: EXTENSIONS AND COMPARATIVE EVALUATION

Kay-See Tan*, University of Pennsylvania
Benjamin C. French, University of Pennsylvania
Andrea B. Troxel, University of Pennsylvania

Conventional longitudinal data analysis methods assume that outcomes are independent of the data-collection schedule. However, the independence assumption may be violated, for example, when adverse events trigger additional physician visits in between prescheduled follow-ups. Observation times may therefore be informative of outcome values, and potentially introduce bias when estimating the effect of covariates on outcomes using a standard longitudinal regression model. Recently developed methods have focused on semi-parametric regression models to account for the relationship between the outcome and observation-time processes, either by specifying an additional regression model for the observation-time process or by conditioning on unobserved latent variables. We extend these methods to accommodate more flexible covariate specifications and to allow adjustment for time-dependent covariates in the observation-time model. We evaluate the statistical properties of these methods under alternative outcome-observation dependence models. In simulation studies, we show that incorrectly specifying the dependence model may yield biased estimates of covariate-outcome associations. We illustrate the implications of different modeling strategies in an application to bladder cancer data. In longitudinal studies with potentially informative observation times, we recommend that analysts carefully explore the dependence mechanism for the outcome and observation-time processes to ensure valid inference regarding covariate-outcome associations.

e-mail: kaystan@mail.med.upenn.edu

107. MULTIVARIATE METHODS

ON SIMPLE TESTS OF DIAGONAL SYMMETRY FOR BIVARIATE DISTRIBUTIONS

Hani M. Samawi*, Georgia Southern University
Robert Vogel, Georgia Southern University

Simple tests of diagonal symmetry for bivariate distributions are proposed. The asymptotic null distributions for all of the proposed tests are derived in this paper. To compare the proposed tests, intensive simulation is conducted to examine the power of the proposed tests under the null and the alternative hypotheses. The proposed tests will be applied to data from previous biomedical studies.

e-mail: hsamawi@georgiasouthern.edu

OPTIMAL DESIGNS FOR BIVARIATE ACCELERATED LIFE TESTING EXPERIMENTS

Xiaojian Xu*, Brock University
Mark Krzeminski, Brock University

Accelerated life testing (ALT) provides a means of obtaining data, on lifetime of highly-reliable products, more quickly by subjecting these products to higher-than-usual levels of stress factors. In this paper, we present the methods for constructing the optimal designs for bivariate ALT in order to estimate percentiles of product life at a normal usage condition when the data observed are time-censored. We assume a Weibull life distribution, and log-linear life-stress relationships with possible heteroscedasticity (non-constant shape parameter) for stress factors. The primary optimality criterion is to minimize the asymptotic variance of the maximum likelihood percentile estimator at the usage stress level combination. For bivariate ALT, this primary criterion yields infinitely-many choices of such designs and in order to further optimally select one of them, we further optimize with respect to a secondary criterion, maximization of either the determinant (D-optimality) or the trace (A-optimality) of Fisher information matrix. Designs are illustrated with practical examples from engineering and a comparison study is presented, which compares results between the two secondary optimality criteria and also between our results and results for homoscedasticity (constant shape parameter).

e-mail: xxu@brocku.ca

JAMES-STEIN TYPE COMPOUND ESTIMATION OF MULTIPLE MEAN RESPONSE FUNCTIONS AND THEIR DERIVATIVES

Limin Feng*, University of Kentucky
Richard Charnigo, University of Kentucky
Cidambi Srinivasan, University of Kentucky

James and Stein proposed an estimator of the mean vector of a p-dimensional multivariate normal distribution in 1961, which produces a smaller risk than the MLE if p >= 3. In this article, we extend their idea to a nonparametric regression setting. More specifically, we present Steinized local regression estimators of p mean response functions and their derivatives. We consider different covariance structures for the error terms, and whether or not a known upper bound for the estimation bias is assumed. We also apply Steinization to compound estimation, another nonparametric regression method which achieves near optimal convergence rates and which is self-consistent: the estimated derivatives equal the derivatives of the estimated mean response functions.

e-mail: limin.feng@uky.edu

BAYESIAN MODELING OF A BIVARIATE DISTRIBUTION WITH CORRELATED CONTINUOUS AND BINARY OUTCOMES

Ross A. Bray*, Baylor University
John W. Seaman Jr., Baylor University
James D. Stamey, Baylor University

We describe a Bayesian model in a simulated clinical trial setting for a bivariate distribution with one binary and one continuous response. The marginal distribution of the binary response is given a Bernoulli distribution with a logit link function and the conditional distribution of the continuous response given the binary response is given a normal distribution with a linear link function. In the simulation, the Bayesian credible sets were obtained through Markov chain Monte Carlo methods using OpenBUGS through R. Parameter estimation is fairly consistent with respect to coverage of the 95% credible sets, however, the posterior estimates of the parameters for the binary response vary more across simulated samples as the probability of the binary response decreases. The marginal posterior variances also increase in the parameters for the binary response as the probability of the binary response decreases, but the marginal posterior variances decrease in the parameters for the conditional continuous response.

e-mail: ross_bray@baylor.edu

ROBUST PARTIAL LEAST SQUARES REGRESSION USING REPEATED MINIMUM COVARIANCE DETERMINANT

Dilrukshika M. Singhabahu*, University of Pittsburgh
Lisa Weissfeld, University of Pittsburgh

Partial Least Squares Regression (PLSR) is often used for high dimensional data analysis where the sample size is limited, the number of variables is large, and when the variables are collinear. PLSR is influenced by outliers and/or influential observations. Since PLSR is based on the covariance matrix of the outcome and the predictor variables, this is a natural starting point for the development of techniques that can be used to identify outliers and to provide stable estimates in the presence of outliers. We focus on the use of the minimum covariance determinant (MCD) method for robust estimation of the covariance matrix when n >> p and modify this method for application to a magnetic resonance imaging (MRI) data set with 1 outcome and 19 predictors. We extend this approach by applying the MCD to generate robust Mahalanobis squared distances (RMSD) in the Y vector and the X matrix separately and detect the outliers and leverage points based on the RMSD. We then remove these observations from the data set and apply PLSR. This approach is applied iteratively until no new outliers and leverage points are detected. Simulation studies demonstrate that PLSR results are improved when using this approach.

e-mail: dms132@pitt.edu

PRINCIPAL COMPONENT ANALYSIS ON HIGH DIMENSIONAL NON-GAUSSIAN DEPENDENT DATA

Fang Han*, Johns Hopkins University
Han Liu, Princeton University

In this paper, we propose a new principal component analysis (PCA) that has the potential to handle large, complex, and noisy datasets. In particular, we study the scenario where the observations are each from a semi-parametric model and drawn from non-i.i.d. processes (m-dependency or a more general phi mixing case). We show that our method can allow weak dependence. In particular, we provide the generalization bounds of convergence for both support recovery and parameter estimation of the proposed method for the non-i.i.d. data. We provide explicit sufficient conditions on the degree of dependence, under which the same parametric rate can be achieved. To our knowledge, this is the first work analyzing the theoretical performance of PCA for the dependent data in high dimensional settings. Our results strictly generalize the analysis in Liu et al. (2012) and the techniques we used have the separate interest for analyzing a variety of other multivariate statistical methods. Our theoretical results are backed up by experiments on synthetic data and real-world genomic and equities data.

e-mail: fhan@jhsph.edu

SPARSE PRINCIPAL COMPONENT REGRESSION

Tamar Sofer*, Harvard School of Public Health
Xihong Lin, Harvard School of Public Health

To account for heterogeneity in study population, in which the covariance between measured outcomes may depend on multiple covariates, we propose to jointly estimate a set of covariance matrices corresponding to multiple subsamples, while burrowing information through a common underlying structure. We define a factor model in which the principal components are linear functions of covariates. Under the assumption that the covariance matrices are sparse, we propose an iterated penalized least squares procedure for estimating model parameters. We use a concave penalty function that satisfies the oracle properties, and show that the estimators of the set of covariance matrices are consistent and sparse when the number of parameters diverges at a sub-exponential rate of the sample size. We propose the Bayesian Information Criterion (BIC) for tuning parameter selection when the number of parameters is moderate and show that it is consistent. To decide if a specific number of estimated principal components well characterizes the covariance matrices, we propose a scree-plot based test for the residual variance. The estimation method is studied by simulations and implemented to study the effect of lifetime smoking on gene methylation in a cohort of elderly men from the Normative Aging Study (NAS).

e-mail: tsofer@hsph.harvard.edu

108. NEW STATISTICAL CHALLENGES FOR LONGITUDINAL/ MULTIVARIATE ANALYSIS WITH MISSING DATA

OUTCOME DEPENDENT SAMPLING FOR CONTINUOUS-RESPONSE LONGITUDINAL DATA

Paul J. Rathouz*, University of Wisconsin, Madison
Jonathan S. Schildcrout, Vanderbilt University School of Medicine
Lee McDaniel, University of Wisconsin, Madison

In outcome dependent sampling (ODS) designs for longitudinal data, the subjects and/or the observations are sampled as a stochastic function of the longitudinal vector of responses. The sampling may for example be a variant on case-control sampling for extreme subjects, or may alternatively sample individual observations from subjects as a function of a surrogate process. ODS results in a type of missingness-by-design, and important questions of optimal design and robust analysis ensue. Several methods have been developed recently for longitudinal binary responses, but less work has been carried out for

the more difficult continuous data case. In this talk, we will explore the use of various approaches to the analysis of continuous or count data under outcome dependent sampling designs.

e-mail: rathouz@biostat.wisc.edu

A SYSTEMATIC APPROACH TO MODEL IGNORABLE MISSINGNESS OF HIGH-DIMENSIONAL DATA

Naisyin Wang*, University of Michigan

We are interested at linking either multi-dimensional or longitudinal covariates to certain outcomes. When the data are not completely observed, a concern could be the potential biases caused by ignoring missing data. We consider a systematic approach that will accommodate multivariate or longitudinal covariates with incomplete data while maintain a high level of feasibility and efficiency. Both theoretical and numerical properties will be discussed.

e-mail: nwangaa@umich.edu

MISSING AT RANDOM AND IGNORABILITY FOR INFERENCES ABOUT INDIVIDUAL PARAMETERS WITH MISSING DATA

Roderick J. Little*, University of Michigan
Sahar Zanganeh, University of Washington

In a landmark paper, Rubin (1976 Biometrika) showed that the missing data mechanism can be ignored for likelihood-based inference about parameters when (a) the missing data are missing at random (MAR), in the sense that missingness does not depend on the missing values after conditioning on the observed data, and (b) distinctness of the parameters of the data model and the missing-data mechanism, that is, there are no a priori ties, via parameter space restrictions or prior distributions, between the parameters of the data model and the parameters of the model for the mechanism. Rubin (1976) described (a) and (b) as the "weakest simple and general conditions under which it is always appropriate to ignore the process that causes missing data". However, it is important to note that these conditions are not necessary for ignoring the mechanism in all situations. We propose conditions for ignoring the missing-data mechanism for likelihood inferences about subsets of the parameters of the data model. We present examples where the missing data are ignorable for some parameters, but the missing data mechanism is missing not at random (MNAR), thus extending the range of circumstances where the missing data mechanism can be ignored.

e-mail: rlittle@umich.edu

SOLVING COMPUTATIONAL CHALLENGES WITH COMPOSITE LIKELIHOOD

Bruce G. Lindsay*, The Pennsylvania State University
Prabhani Kuruppumullage, The Pennsylvania State University

We build a mixture type model for the purpose of two way clustering of data. The resulting likelihood requires calculations that grow at an exponential rate in the data dimensions. The exact calculation of this likelihood is infeasible in the examples we consider. As an alternative we have constructed a composite likelihood whose calculation effort is similar to a standard mixture model, growing linearly in the data dimensions. It provides promising results. We then consider how to replace the standard likelihood tools in this setting. For example, we develop an algorithm to replace the standard mixture EM, we develop a replacement to the standard way of clustering points, and we develop a composite likelihood to use with missing data. We use the latter to perform a version of likelihood cross-validation for assessing the number of row and column clusters.

e-mail: bgl@psu.edu

109. STATISTICAL INFORMATION INTEGRATION OF -OMICS DATA

THE INFERENCE OF DRUG PATHWAY ASSOCIATIONS THROUGH JOINT ANALYSIS OF DIVERSE HIGH THROUGHPUT DATA SETS

Haisu Ma, Yale University
Ning Sun, Yale University
Hongyu Zhao*, Yale University

Pathway-based drug discovery considers the therapeutic effects of compounds in the global physiological environment. Because the target pathways and mechanism of action for many compounds are still unknown, and there are also some unexpected off-target effects, the inference of drug-pathway associations is a crucial step to fully realize the potential of system-based pharmacological research. Transcriptome data offer valuable information on drug pathway targets because the pathway activities may be reflected through gene expression levels. In this talk, we will introduce a Bayesian sparse factor analysis model to jointly analyze the paired gene expression and drug sensitivity datasets measured across the same panel of samples. The model enables direct incorporation of prior knowledge regarding gene-pathway and/or drug-pathway associations to aid the discovery of new association relationships. Our results suggest that is a promising approach for the identification of drug targets. This model also provides a general statistical framework for pathway-based integrative analysis of other types of -omics data.

e-mail: hongyu.zhao@yale.edu

iASeq: INTEGRATIVE ANALYSIS OF ALLELE-SPECIFICITY OF PROTEIN-DNA INTERACTIONS IN MULTIPLE CHIP-SEQ DATASETS

Yingying Wei, Johns Hopkins Bloomberg School of Public Health
Xia Li, Chinese Academy of Sciences
Qianfei Wang, Chinese Academy of Sciences
Hongkai Ji*, Johns Hopkins Bloomberg School of Public Health

ChIP-seq provides new opportunities to study allele-specific protein-DNA binding (ASB). Currently, little is known about the correlation patterns of allele-specificity among different transcription factors and epigenetic marks. Moreover, detecting allelic imbalance from a single ChIP-seq dataset often has low statistical power since only sequence reads mapped to heterozygote SNPs are informative for discriminating two alleles. We develop a new method iASeq to address both issues by jointly analyzing multiple ChIP-seq datasets. iASeq uses a Bayesian hierarchical mixture model to identify correlation patterns of allele-specificity among multiple proteins. Using the discovered correlation patterns, the model allows one to borrow information across datasets to improve detection of allelic imbalance. Application of iASeq to 106 ChIP-seq samples from 40 ENCODE datasets in GM12878 cells reveals that allele-specificity of multiple proteins are highly correlated. The analysis also demonstrates the ability of iASeq to improve allelic inference compared to analyzing each individual dataset separately. iASeq illustrates the value of integrating multiple datasets in the allele-specificity inference and offers a new tool to better analyze ASB.

e-mail: hji@jhsph.edu

INTEGRATIVE ANALYSIS OF RNA AND DNA SEQUENCING DATA IDENTIFIES TISSUE SPECIFIC TRANSCRIPTOMIC SIGNATURES OF EVOKED INFLAMMATION IN HUMANS

Minyao Li*, University of Pennsylvania

Recent genome wide association studies have provided increased insight into the genetic basis of complex cardio-metabolic diseases including type 2 diabetes and atherosclerotic cardiovascular diseases. Such discoveries, however, only explain a small proportion of the heritability suggesting genetic influences other than common DNA variation need to be considered. Knowledge of the transcriptome is essential for a more complete understanding of the inherited functional elements of the genome. The advent of high-throughput sequencing has demonstrated the existence of a far greater transcriptomic complexity and diversity than previously catalogued. Here, we present an integrative analysis of RNA and DNA sequencing data to identify novel tissue specific transcriptomic signatures of evoked inflammation in healthy human subjects. We

show that an integrative analysis combined with human experimental models can reveal biologically relevant transcriptomic changes modulated by evoked inflammation, including differentially expressed isoforms, alternative splicing events, allelic imbalance and RNA editing.

e-mail: mingyao@mail.med.upenn.edu

110. EXPLORING INTERACTIONS IN BIG DATA

SPECTRAL METHODS FOR ANALYZING BIG NETWORK DATA

Jiashun Jin*, Carnegie Mellon University

We propose a new method for network community detection. We approach this problem with the recent Degree Corrected Block Model (DCBM), and using the leading eigenvectors of the adjacency matrix for clustering. The method was successfully applied to the karate data and the web blogs data, with error rates 1/34 and 58/1222, respectively. The method is easy to use, computationally fast, and compare much more satisfactory with ordinary spectral methods.

e-mail: jiashun@stat.cmu.edu

LINK PREDICTION FOR PARTIALLY OBSERVED NETWORKS

Yunpeng Zhao, George Mason University
Elizaveta Levina*, University of Michigan
Ji Zhu, University of Michigan

Link prediction is one of the fundamental problems in network analysis. In many applications, notably in genetics, a partially observed network may not contain any negative examples of absent edges, which creates a difficulty for many existing supervised learning approaches. We present a new method which treats the observed network as a sample of the true network with different sampling rates for positive and negative examples, where the sampling rate for negative examples is allowed to be 0. The sampling rate itself cannot be estimated in this setting, but a relative ranking of potential links by their probabilities can be, which is sufficient in many applications. To obtain these rankings, we utilize information on node covariates as well as on network topology, and set up the problem as a loss function measuring the fit plus a penalty measuring similarity between nodes. Empirically, the method performs well under many settings, including when the observed network is sparse and when the sampling rate is low even for positive examples. We illustrate the performance of our method and compare to others on an application to a protein-protein interaction network.

e-mail: elevina@umich.edu

INTERACTION SELECTION FOR ULTRA HIGH-DIMENSIONAL DATA

Hao Zhang*, University of Arizona
Ning Hao, University of Arizona

For the ultra-high dimensional data, it is extremely challenging to identify important interaction effects among covariates. The first major challenge is implementation feasibility. When the data dimension is more than hundreds of thousands, the total number of interactions is enormous and far beyond capacity of standard software and computers. The second difficulty is the computation speed, even if doable, required to solve big-scaled optimization problems. Asymptotic theory poses additional key challenges. We propose a new class of methodologies, along with efficient computational algorithms, to tackle these issues. The new methods are featured with feasible implementation, fast speed, and desired theoretical properties. Various examples are presented to illustrate the new proposals.

e-mail: hzhang.work@gmail.com

CONSISTENT CROSS-VALIDATION FOR TUNING PARAMETER SELECTION IN HIGH-DIMENSIONAL VARIABLE SELECTION

Yang Feng*, Columbia University
Yi Yu, Fudan University

For variable selection in high dimensional setting, we systematically investigate the properties of several cross-validation methods for selecting the penalty parameter in the popular penalized maximum likelihood method. We show that the popular leave-one-out cross-validation and K-fold cross-validation (with any pre-specified value of K) are both inconsistent in terms of model selection. A new cross-validation procedure, Consistent Cross-Validation (CCV) is proposed. Under certain technical conditions, CCV is shown to enjoy the model selection consistency property. Extensive simulations and real data analysis are conducted, supporting the theoretical results.

e-mail: yangfeng@stat.columbia.edu

111. ASSESSING THE CLINICAL UTILITY OF BIOMARKERS AND STATISTICAL RISK MODELS

A NEW FRAMEWORK FOR ASSESSING THE RISK STRATIFICATION OF MARKERS AND STATISTICAL RISK MODELS

Hormuzd Katki*, National Cancer Institute, National Institutes of Health

The risk stratification of markers or models is usually assessed with measures of discrimination. However, there is substantial controversy about how to use measures of discrimination for assessing the risk stratification or clinical utility of markers or models. Instead, I propose a new framework for assessing risk stratification based on the absolute change in the absolute risk of disease identified by using the marker or model. This measure has a clear and useful clinical interpretation: how much a patient's absolute risk of disease will change by using the marker or model to make clinical decisions. The key quantities in the framework have natural clinical interpretations from the point of view of absolute risk differences (or its reciprocal, the number needed to treat) and marker/model positivity. While measures of discrimination have the Achilles' heel of weighing false-positive and false-negative results equally, these new measures have the Achilles' heel of requiring that each patient rationally manages his risk by 'managing equal risks equally'. I show examples of the usefulness of these new measures in my experience serving on the cervical cancer screening guidelines committee. I demonstrate how these new measures shed useful light on controversial questions about the value of human papillomavirus (HPV) testing for triaging abnormal Pap smear findings.

e-mail: katkih@mail.nih.gov

INCORPORATING COVARIATES IN ASSESSING THE PERFORMANCE OF MARKERS FOR TREATMENT SELECTION

Holly Janes*, Fred Hutchinson Cancer Research Center

Biomarkers that predict the efficacy of a treatment are highly sought after in many clinical contexts, especially where the treatment is sufficiently toxic or costly so that its provision is only warranted if the expected benefit of treatment is suitably large. When evaluating the performance of a candidate treatment selection marker, there are often additional variables that should be considered. These might include factors that affect the marker measurement but that are independent of treatment effect, such as assaying laboratory; other markers that predict treatment effect; and factors that might affect the performance of the marker, such as patient characteristics. We describe conceptual approaches to accommodating variables of each type. Our focus is on measuring the performance of the marker by considering the effect of marker-based treatment on the expected rate of adverse clinical events. Methods for estimation and inference are described and illustrated in the estrogen-receptor positive breast cancer treatment context where the task is to identify women who benefit from adjuvant chemotherapy. In this setting, patient clinical characteristics are important treatment selection markers in their own right, and these variables may also affect the performance of novel candidate markers.

e-mail: hjanes@fhcrc.org

PERSONALIZED EVALUATION OF BIOMARKER VALUE: A COST-BENEFIT PERSPECTIVE

Ying Huang*, Fred Hutchinson Cancer Research Center

A biomarker or medical test that has a potential to inform treatment decisions in clinical practice may be costly to measure. Understanding the extra benefit provided by the marker is therefore important to patients and clinicians who are making the decisions about whether to have a patient's biomarker measured. Common methods for evaluating a biomarker's utility in a general population are not ideal for this purpose: a biomarker that is useful for guiding treatment decisions to the general population will have different values to different patients due to the individual differences in their response to treatment and in their tolerance of the disease harm and the treatment cost. In this talk, we propose a new tool to quantify a biomarker's treatment-selection value to individual patients, which integrates two pieces of personal information including a patient's baseline risk factors and the patient's input about the ratio of treatment cost relative to disease cost. We develop estimation methods for both randomized trials and cohort studies.

e-mail: yhuang@fhcrc.org

112. DESIGN OF CLINICAL TRIALS FOR TIME-TO-EVENT DATA

BAYESIAN SEQUENTIAL META-ANALYSIS DESIGN IN EVALUATING CARDIOVASCULAR RISK IN A NEW ANTIDIABETIC DRUG DEVELOPMENT PROGRAM

Joseph G. Ibrahim*, University of North Carolina, Chapel Hill
Ming-Hui Chen, University of Connecticut
Amy Xia, Amgen Inc.
Thomas Liu, Amgen Inc.
Violeta Hennessey, Amgen Inc.

Recently, the Center for Drug Evaluation and Research at the Food and Drug Administration (FDA) released a guidance document that makes recommendations about how to demonstrate that a new anti-diabetic therapy to treat type 2 diabetes is not associated with an unacceptable increase in cardiovascular risk. One of the recommendations from the guidance is that phase 2 and 3 trials should be appropriately designed and conducted so that a meta-analysis can be performed; the phase 2 and 3 programs should include patients at higher risk of cardiovascular events; and it is likely that the controlled trials will need to last more than the typical 3 to 6 months duration to obtain enough events and to provide data on longer-term cardiovascular risk (e.g., minimum 2 years) for these chronically used therapies. In this context, we develop a new Bayesian sequential meta-analysis approach using survival regression models to assess whether the size of a clinical development program is adequate to evaluate a particular safety endpoint. We propose a Bayesian sample size determination methodology for sequential meta-analysis clinical trial design with a focus on controlling the family-wise type I error and power. The proposed methodology is applied to the design of a new anti-diabetic drug development program for evaluating cardiovascular risk.

e-mail: ibrahim@bios.unc.edu

USING DATA AUGMENTATION TO FACILITATE CONDUCT OF PHASE I/II CLINICAL TRIALS WITH DELAYED OUTCOMES

Ying Yuan*, University of Texas MD Anderson Cancer Center
Ick Hoon Jin, University of Texas MD Anderson Cancer Center
Peter Thall, University of Texas MD Anderson Cancer Center

Phase I/II clinical trial designs combine conventional phase I and phase II trials by using both toxicity and efficacy to determine an optimal dose of a new agent. While many phase I/II designs have been proposed, they have seen very limited use. A major practical impediment, for phase I/II and many other adaptive clinical trial designs, is that outcomes used by adaptive decision rules must be observed soon after the start of therapy in order to apply the rules to choose treatments or doses for new patients. In phase I/II, a severe logistical problem occurs if either toxicity or efficacy cannot be scored quickly, for example if either outcome takes up to six weeks to evaluate but two or more patients are accrued per month. We propose a general methodology for this problem that treats late-onset outcomes as missing data. Given a probability model for the times to toxicity and efficacy as functions of dose, we use data augmentation to impute missing binary outcomes from their posterior predictive distributions based on both partial follow-up information and complete outcome data. Using the completed data, we apply the phase I/II design's decision rules, subject to dose safety and efficacy admissibility requirements. We illustrate the method with two cancer clinical trials, including computer stimulations.

e-mail: yyuan@mdanderson.org

BAYESIAN DESIGN OF SUPERIORITY CLINICAL TRIALS FOR RECURRENT EVENTS DATA WITH APPLICATIONS TO BLEEDING AND TRANSFUSION EVENTS IN MYELODYPLASTIC SYNDROME

Ming-Hui Chen*, University of Connecticut
Joseph G. Ibrahim, University of North Carolina, Chapel Hill
Donglin Zeng, University of North Carolina, Chapel Hill
Kuolung Hu, Amgen Inc.
Catherine Jia, Amgen Inc.

In many biomedical studies, patients may experience the same type of recurrent event repeatedly over time, such as bleeding, multiple infections and disease. In this paper, we aim to design a pivotal clinical trial in which lower risk Myelodysplastic syndrome (MDS) patients are treated with MDS disease modifying therapies. One of the key study objectives is to demonstrate the investigational product (treatment) effect on reduction of platelet transfusion and bleeding events while receiving MDS therapies. In this context, we propose a new Bayesian approach for the design of superiority clinical trials using recurrent events regression models. The recurrent events data from a completed phase 2 trial is incorporated into the Bayesian design via the power prior of Ibrahim and Chen (2000). An efficient MCMC sampling algorithm, a predictive data generation algorithm, and a simulation-based algorithm are developed for sampling from the fitting posterior distribution, generating the predictive recurrent events data, and computing various design quantities such as type I error and power, respectively. Various properties of the proposed methodology are examined and an extensive simulation study is conducted.

e-mail: ming-hui.chen@uconn.edu

113. STATISTICAL ANALYSIS OF SUBSTANCE ABUSE DATA

TIME-VARYING COEFFICIENT MODELS FOR LONGITUDINAL MIXED RESPONSES

Esra Kurum, Istanbul Medeniyet University, Istanbul, Turkey
Runze Li*, The Pennsylvania State University
Saul Shiffman, University of Pittsburgh
Weixin Yao, Kansas State University

Motivated by an empirical analysis of ecological momentary assessment data (EMA) collected in a smoking cessation study, we propose a joint modeling technique for estimating the time-varying association between two intensively measured longitudinal responses: a continuous one and a binary one. A major challenge in joint modeling these responses is the lack of a multivariate distribution. We suggest introducing a normal latent variable underlying the binary response and factorizing

the model into two components: a marginal model for the continuous response, and a conditional model for the binary response given the continuous response. We develop a two-stage estimation procedure and establish the asymptotic normality of the resulting estimators. We also derived the standard error formulas for estimated coefficients. We conduct a Monte Carlo simulation study to assess the finite sample performance of our procedure. The proposed method is illustrated by an empirical analysis of smoking cessation data, in which the important question of interest is to investigate the association between urge to smoke, continuous response, and the status of alcohol use, the binary response, and how this association varies over time.

e-mail: rli@stat.psu.edu

METHODS FOR MEDIATION ANALYSIS IN ALCOHOL INTERVENTION STUDIES

Joseph W. Hogan*, Brown University
Michael J. Daniels, University of Texas, Austin

Studies of alcohol intervention are designed to stop or reduce alcohol intake. In this talk, we provide an overview of statistical methods for assessing mediation effects. We present a unified method for handling one mediator or several correlated mediators. We argue that natural direct and indirect effects are the most relevant summaries of mediation effects, because most mediators in behavioral intervention studies cannot be externally manipulated. An important feature of mediation analyses is that the joint distribution of relevant potential outcomes cannot be identified by observed data; we describe a sensitivity analysis framework to address this. Our methods are illustrated on a study of a behavioral intervention to reduce alcohol consumption among HIV-infected individuals.

e-mail: jhogan@stat.brown.edu

MIXTURE MODELS FOR THE ANALYSIS OF DRINKING DATA IN CLINICAL TRIALS IN ALCOHOL DEPENDENCE

Ralitza Gueorguieva*, Yale University

Some of the heterogeneity of clinical findings in studies evaluating treatment efficacy for alcohol dependence can be attributed to the wide use of standard statistical analytical tools of summary drinking measures that poorly reflect the distributions, variability, and complexity of drinking data. Mixture models provide tools for more appropriate handling both the distribution of drinking data and variability in treatment response. In particular, mixture distributions allow for more accurate description of drinking data with excess zeroes and over-dispersion. Growth mixture models (GMM) allow assessment of population heterogeneity in treatment response over time. However, in GMM selection of number of classes is challenging and often based on combination of statistical criteria and subject-matter interpretation. We will present mixture models for daily drinking data and illustrate the flexibility and challenges of such models on data from two large clinical trials in alcohol dependence. Parameter estimation, statistical inference, assessment of model fit and interpretation will be discussed.

e-mail: ralitza.gueorguieva@yale.edu

ANALYZING REPEATED MEASURES SEMI-CONTINUOUS DATA, WITH APPLICATION TO AN ALCOHOL DEPENDENCE STUDY

Lei Liu*, Northwestern University
Robert Strawderman, University of Rochester
Bankole Johnson, University of Virginia
John O'Quigley, The´orique et Applique´e Universite´ Pierre et Marie Curie, Paris

Two-part random effects models have been applied to repeated measures of semi-continuous data, characterized by a mixture of a substantial proportion of zero values and a skewed distribution of positive values. In the original formulation of this model, the natural logarithm of the positive values is assumed to follow a normal distribution with a constant variance parameter. In this article, we review and consider three extensions of this model, allowing the positive values to follow (a) a generalized gamma distribution, (b) a log-skew-normal distribution, and (c) a normal distribution after the Box-Cox transformation. We allow for the possibility of heteroscedasticity. Maximum likelihood estimation is shown to be conveniently implemented in SAS Proc NLMIXED. The performance of the methods is compared through applications to daily drinking records in a secondary data analysis from a randomized controlled trial of topiramate for alcohol dependence treatment. We find that all three models provide a significantly better fit than the log-normal model, and there exists strong evidence for heteroscedasticity. We also compare the three models by the likelihood ratio tests for nonnested hypotheses. The results suggest that the generalized gamma distribution provides the best fit, though no statistically significant differences are found in pairwise model comparisons.

e-mail: lei.liu@northwestern.edu

114. GLM AND BEYOND: BOOK AUTHORS DISCUSS CUTTING-EDGE APPROACHES AND THEIR CHOSEN VENUE FOR PUBLICATION

FAME AND FORTUNE IN GLM-RELATED TEXTBOOKS

James W. Hardin*, University of South Carolina

I will discuss the nature of the information included in those GLM-related textbooks on which I am a coauthor, illustrated software in those texts, how (and by whom) decisions are reached about the extent to which software is illustrated, support (and non-support) given by publishers, required formats for production, author responsibilities, and professional credit received. In addition, I will discuss financial arrangements, and what reasonable amounts prospective authors might expect from royalties.

e-mail: jhardin@sc.edu

DECAYING PRODUCT STRUCTURES IN EXTENDED GEE (QLS) ANALYSIS OF LONGITUDINAL AND DISCRETE DATA

Justine Shults*, University of Pennsylvania Pereleman School of Medicine

Longitudinal studies such as clinical trials often involve discrete outcomes that are unbalanced with respect to the number of measurements per patient, are unequally spaced in time, and have non-constant variance. I use quasi-least squares (QLS) to implement decaying product correlation structures with scalar parameters that vary over time, which allows for appropriate and robust analysis of data with these complex features. These results will be described in Quasi-Least Squares Regression- Extended Generalized Estimating Equation Methodology (CRC Press, 2013) and/or Logistic Regression for Correlated Binary Data (Springer, 2013). I also describe my experience in co-authoring texts that summarize and extend my former work in this area; this should hopefully be of interest to other biostatisticians who are considering writing a text.

e-mail: jshults@mail.med.upenn.edu

115. MORE METHODS FOR HIGH-DIMENSIONAL DATA ANALYSIS

RANDOM FORESTS FOR CORRELATED PREDICTORS IN IMMUNOLOGIC DATA

Joy Toyama*, University of California, Los Angeles
Christina Kitchen, University of California, Los Angeles

Determining associations among clinical phenotypes used for HIV therapeutics is a difficult problem to surmount due to the large number of potential predictors of the related immunophenotypic variables compared to the number of observations. Further exacerbating this problem is the fact that the predictor variables are highly correlated. Ensemble classifiers have been developed and used to handle such data. Random forests is a well-known nonparametric method which can illustrate complex interactions between the immunophenotpyic variables. In general, random forests present more stable results than tree classifiers alone and other ensemble methods. While this key characteristic holds true, it has been shown that variable importance measures and can lead to biased results when there is high correlation

among the potential predictors. When the goal is generating stable lists of important variables as opposed to prediction, allowing correlated variables to be selected is important. We propose a modification that includes more randomness to break this correlation and allow random forests to choose correlated variables more often and thus improve accuracy of the variance importance metrics.

e-mail: jtoyama@ucla.edu

EVALUATING GENE-GENE AND GENE-ENVIRONMENTAL INTERACTIONS USING THE TWO-STAGE RF-MARS APPROACH: THE LUNG CANCER EXAMPLE

Hui-Yi Lin*, Moffitt Cancer Center & Research Institute

The major of current genetic variation studies only evaluate gene-environmental (GE) interactions for single nucleotide polymorphisms (SNPs) with a significant main effect. Pure interactions with weak or no main SNP effects will be neglected. In this study, we explored gene-gene (GG) and GE interactions associated with lung cancer risk using the two-stage Random Forests plus Multivariate Adaptive Regression Splines (RF-MARS) approach. Including smoking pack year, a strong predictor (p=1.5*10-94), in a model may disguise the true associations, so our analyses were stratified by pack-year status [light (<=30) and heavy smokers (>30 pack-years)]. We evaluated 38 SNPs in the 10 candidate genes using 1764 non-small cell lung carcinoma cases and 2106 controls in the GENEVA/GEI lung cancer and smoking study. Only ever smokers were included. In the model of light smokers, those with the GG genotype of rs12441998 in CHRNB4 and 15-30 pack-year had a higher risk than other light smokers (OR=3.6, 95% CI=2.8-4.6, p=4.5*10-22). The interaction of rs578776 in CHRNA3 and pack-year was also significant (p=0.003). For the heavy smoker model, one GE interaction (rs4646421 * pack-year, p=0.015), one main effect of rs1051730 in CHRNA3 (p=1.8*10-6) and two GG interactions (rs1051730* rs4975605 and rs1051730 * rs11636753) were significantly associated with lung cancer risk.

e-mail: hui-yi.lin@moffitt.org

STATISTICAL LINKAGE ACROSS HIGH DIMENSIONAL OBSERVATIONAL DOMAINS

Leonard Hearne*, University of Missouri
Toni Kazic, University of Missouri

It is reasonably common in many experimental sciences to have high-dimensional data from multiple observational domains collected from the same experimental units. These domains may be genotypic, phenotypic, proteomic, or any other domain where distinct aspects of experimental units are being measured. When we are interested in comparing relationships between homogeneous regions in one high dimensional domain with one or more regions in another high dimensional domain, the number of possible comparisons may be extremely large and not known a priori. We present an outline of procedures for identifying possible relationships or inferential links between regions in one domain and regions in another domain. If the data are dense enough, then statistical mesures of association can be computed. Though these procedures are extremely compute intensive, they provide a measure of the probability of inter-observational domain associations.

e-mail: hearnel@missouri.edu

HYPOTHESIS TESTING USING A FLEXIBLE ADDITIVE LEAST SQUARES KERNEL MACHINE APPROACH

Jennifer J. Clark*, University of North Carolina, Chapel Hill
Mike Wu, University of North Carolina, Chapel Hill
Arnab Maity, North Carolina State University

High throughput biotechnology has led to a revolution within modern biomedical research. New omic studies offer to provide researchers with intimate understanding of how genetic features (SNPs, genes, proteins, etc.) influence disease outcomes and hold the potential to comprehensively address key biological, medical, and public health questions. However, the high-dimensionality of the data, the limited availability of samples, and poor understanding of how genomic features influence the outcome pose a grand challenge for statisticians. The field keenly needs powerful new statistical methods that can accommodate complex, high dimensional data. Multi-feature testing has proven to be a useful strategy for analysis of genomic data. Existing methods generally require adjusting for covariates in a simplistic, linear fashion. We propose to model the genomic features and the complex covariates using the flexible additive least square kernel machine (ALSKM) regression framework. We demonstrate via simulations and real data applications that our approach allows for accurate modeling of covariates and subsequently improved power to detect true multi-feature effects and better control of type I error when covariate effects are complex, while losing little power when covariates are relatively simple.

e-mail: jjclark@live.unc.edu

NODE- AND GRAPH-LEVEL PROPERTIES OF CENTRALITY IN SMALL NETWORKS

C. Christina Mehta*, Emory University
Vicki S. Hertzberg, Emory University

Although much applied research has been conducted describing node- and graph-level characteristics of real networks, there is less information about their properties. We describe node- and graph-level properties of 3 commonly used network measures, relative maximum centrality (node-level) and relative centralization (graph-level) for closeness, betweenness, and degree, the relationship between them, and their variation by network size. We developed a new method to create simple graphs encompassing the full range of centralization values, then examined the relationship between maximum centrality and centralization for 5-14 node graphs. The observed value range indicates that we can successfully create graphs that span most of the theoretical range. Results suggest that highly centralized graphs are infrequent while moderately centralized graphs are most common for all 3 measures. A Pearson correlation showed increasing positive correlation by size for graphs without a connectedness restriction. These results provide insight into the relative frequency of obtaining a centralization value and the range of centralization values for a maximum centrality value.

e-mail: christina.mehta@emory.edu

ADAPTIVE NONPARAMETRIC EMPIRICAL BAYES ESTIMATION VIA WAVELET SERIES

Rida Benhaddou*, University of Central Florida

Empirical Bayes (EB) methods are estimation techniques in which the prior distribution is estimated from the data. They provide powerful tools for data analysis in biomedical sciences with applications ranging from studies of sequencing-based transcriptional profiling in transcriptome analysis to metabolite identification in gas chromatography mass spectrometry. We propose generalization of the linear empirical Bayes estimation method which takes advantage of the exibility of the wavelet techniques. We present an empirical Bayes estimator as a wavelet series expansion and estimate coefficients by minimizing the prior risk of the estimator. Although wavelet series have been used previously for EB estimation, the method suggested in the paper is completely novel since the EB estimator as a whole is represented as a wavelet series rather than its components. Moreover, the method exploits de-correlating property of wavelets which is not instrumental for the former wavelet-based EB techniques. We also proposes an adaptive choice of the resolution level using Lepskii (1997) method. The method is computationally efficient and provides asymptotically optimal EB estimators posterior risks of which tend to zero at an optimal rate. The theory is supplemented by numerous examples.

e-mail: ridab2009@knights.ucf.edu

116. SEMIPARAMETRIC AND NONPARAMETRIC MODELS FOR LONGITUDINAL DATA

ADJUSTMENT OF DEPENDENT TRUNCATION WITH INVERSE PROBABILITY OF WEIGHTING

Jing Qian*, University of Massachusetts, Amherst
Rebecca A. Betensky, Harvard School of Public Health

Increasing number of clinical trials and observational studies are conducted under complex sampling involving truncation. Ignoring the issue of truncation or incorrectly assuming quasi-independence can lead to bias and incorrect results. Currently available approaches for dependently truncated data are sparse. We present an inverse probability weighting method for estimating the survival function of a failure time subject to left truncation and right censoring. Our method allows adjustment for informative truncation due to factors affecting both time-to-event and truncation. Both inverse probability of truncation weighted product-limit estimator and Cox partial likelihood estimators are developed. Simulation studies show that the proposed method performs well in finite sample. We illustrate our approach in a real data application.

e-mail: qian@schoolph.umass.edu

NONPARAMETRIC INFERENCE FOR INVERSE PROBABILITY WEIGHTED ESTIMATORS WITH A RANDOMLY TRUNCATED SAMPLE

Xu Zhang*, University of Mississippi Medical Center

A randomly truncated sample appears when the independent variables T and L are observable if $L < T$. The truncated version Kaplan-Meier estimator is known to be the standard estimation method for the marginal distribution of T or L. The inverse probability weighted (IPW) estimator was suggested as an alternative and its agreement to the truncated version Kaplan-Meier estimator has been proved. This paper centers on the weak convergence of IPW estimators and variance decomposition. The paper shows that the asymptotic variance of an IPW estimator can be decomposed into two sources. The variation for the IPW estimator using known weight functions is the primary source, and the variation due to estimated weights should be included as well. Variance decomposition establishes the connection between a truncated sample and a biased sample with know probabilities of selection. A simulation study was conducted to investigate the practical performance of the proposed variance estimators, as well as the relative magnitude of two sources of variation for various truncation rates. A blood transfusion infected AIDS data set is analyzed to illustrate the nonparametric inference discussed in the paper.

e-mail: xzhang2@umc.edu

EM ALGORITHM FOR REGRESSION ANALYSIS OF INTERVAL-CENSORED DATA UNDER THE PROPORTIONAL HAZARDS MODEL

Lianming Wang*, University of South Carolina
Christopher S. McMahan, Clemson University
Michael G. Hudgens, University of North Carolina, Chapel Hill
Xiaoyan Lin, University of South Carolina

The proportional hazards model (PH) is the probably the most popular semiparametric regression model for analyzing time-to-event data. In this paper, we propose a novel frequentist approach using EM algorithm to analyze interval-censored data under the PH model. First, we model the cumulative baseline hazard function with an additive form of monotone splines. Then we expand the observed likelihood to augmented data likelihood as a product of Poission probability mass functions by introducing two sets of Poisson latent variables. Derived based on the augmented likelihood, our EM algorithm involves simply solving a system of equations for the regression parameters and updating the spline coefficients in closed form iteratively. In addition, our EM algorithm provides variance for all parameter estimates in closed form. Simulation study shows that our method works well and converges fast. We illustrate our method to a clinical trial data about mother-to-child transmission of HIV.

e-mail: wang99@mailbox.sc.edu

INFERENCE ON CONDITIONAL QUANTILE RESIDUAL LIFE FOR CENSORED SURVIVAL DATA

Wen-Chi Wu*, University of Pittsburgh
Jong-Hyeon Jeong, University of Pittsburgh

The quantile residual life function at time t has been conducted in the univariate settings with or without covariates. However, patients may experience two different types of events and the conditional quantile residual lifetimes to a later event after experiencing the first event might be of utmost interest, which requires a bivariate modeling of quantile residual lifetimes subject to right censoring. Under a bivariate setting, the conditional survival function given one of random variables can be estimated by the Kaplan-Meier estimator with a kernel density estimator and bandwidth which provides different weights for censored and uncensored data points. In this study, a nonparametric estimator for conditional quantile residual life function at a fixed time point is proposed by inverting a function of Kaplan-Meier estimators. The estimator is shown to be asymptotically consistent and converges to a zero-mean Gaussian process. A test statistic for comparing the ratio of two conditional quantile residual lifetimes is performed. Numerical studies validate the proposed test statistic in terms of type I error probabilities and demonstrate that the estimator performs well for a moderate sample size. The proposed method is applied to a breast cancer dataset from a phase III clinical trial.

e-mail: wew25@pitt.edu

REGRESSION ANALYSIS OF BIVARIATE CURRENT STATUS DATA UNDER THE GAMMA-FRAILTY PROPORTIONAL HAZARDS MODEL USING EM ALGORITHM

Naichen Wang*, University of South Carolina
Lianming Wang, University of South Carolina
Christopher S. McManhan, Clemson University

The Gamma-frailty proportional hazards model is widely used to model correlated survival times. In this paper, we propose a novel EM algorithm to analyze bivariate current status data, in which the two failure times are correlated and both have current status data structure. The fundamental steps for constructing our EM algorithm include the use of monotone splines for the conditional baseline hazard functions and the introduction of Poisson latent variables. Our EM algorithm involves solving a system of equations for the regression parameters and gamma variance parameter and updating the spline coefficients in closed form. The EM algorithm is robust to initial values and converges fast. In addition, the variance estimates are provided in closed form. Our method is evaluated by a simulation study and is illustrated by a real life data set about hepatitis B and HIV among the population of prisoners in the Republic of Ireland.

e-mail: wangn@email.sc.edu

NONPARAMETRIC RESTRICTED MEAN ANALYSIS ACROSS MULTIPLE FOLLOW-UP INTERVALS

Nabihah Tayob*, University of Michigan
Susan Murray, University of Michigan

This research provides a nonparametric estimate of tau-restricted mean survival that uses additional follow-up information beyond tau, when appropriate, to improve precision. The tau-restricted mean residual life function and its associated confidence bands are a tool to assess the stability of disease prognosis and the validity of combining follow-up intervals for this purpose. The variance of our estimate must account for correlation between incorporated follow-up windows and we follow an approach by Woodruff (1971) that linearizes random components of the estimate to simplify calculations. Both asymptotic closed form calculations and simulation studies recommend selection of follow-up intervals spaced approximately tau/2 apart. In simulations, the variance we propose performs better than the standard sandwich variance estimate. Our analysis approach is illustrated in two settings summarizing prognosis of idiopathic pulmonary fibrosis patients and asprin treated diabetic retinopathy patients who had deferred photocoagulation.

e-mail: tayob@umich.edu

A PIECEWISE LINEAR CONDITIONAL SURVIVAL FUNCTION ESTIMATOR

Seung Jun Shin*, North Carolina State University
Helen Zhang, North Carolina State University
Yichao Wu, North Carolina State University

In survival data analysis, the conditional survival function (CSF) is often of interest in order to study the relationship between a survival time and explanatory covariates. In this article, we propose a piecewise linear conditional survival function estimator based on the complete solution surface we develop for the censored kernel quantile regression (CKQR). The proposed CSF estimator is a flexible nonparametric method which does not require any specific model assumption such as linearity of the survival time and proportional hazards. One important advantage of the estimator is that it can handle very high-dimensional covariates. We carry out asymptotic analysis to justify the estimator theoretically and numerical experiments to demonstrate its competitive finite-sample performance under various scenarios.

e-mail: sshin@ncsu.edu

117. TIME SERIES ANALYSIS

A PRIOR FOR PARTIAL AUTOCORRELATION SELECTION

Jeremy Gaskins*, University of Florida
Michael Daniels, University of Texas, Austin

Modeling a correlation matrix can be a difficult statistical task due to the positive definiteness and unit diagonal constraints. Because the number of parameters increases quadratically in the dimension, it is often useful to consider a sparse parameterization. We introduce a prior distribution on the set of correlation matrices through the set of partial autocorrelations (PACs), each of which vary independently over [-1,1]. The prior for each PAC is a mixture of a zero point mass and a continuous component, allowing for a sparse representation. The structure implied under our prior is readily interpretable because each zero PAC implies a conditional independence relationship in the distribution of the data. The PAC priors are compared to standard methods through a simulation study. Further, we demonstrate our priors in a multivariate probit analysis on data from a smoking cessation clinical trial.

e-mail: jgaskins@stat.ufl.edu

BAYESIAN ANALYSIS OF TIME-SERIES DATA UNDER CASE-CROSSOVER DESIGNS: POSTERIOR EQUIVALENCE AND INFERENCE

Shi Li*, University of Michigan
Bhramar Mukherjee, University of Michigan
Stuart Batterman, University of Michigan
Malay Ghosh, University of Florida

Case-crossover designs are widely used to study short-term exposure effects on the risk of acute adverse health events. The contribution of this paper is two-fold. First, the paper establishes Bayesian equivalence results that require characterization of the set of priors under which the posterior distributions of the risk ratio parameters based on a case-crossover and time-series analysis are identical. Second, the paper studies inferential issues under case-crossover designs in a Bayesian framework. Traditionally, a conditional logistic regression is used for inference on risk-ratio parameters in case-crossover studies. We consider instead a more general full likelihood-based approach which makes less restrictive assumptions on the risk functions. We propose a semi-parametric Bayesian approach using a Dirichlet process prior to handle the random nuisance parameters that appear in a full likelihood formulation. We carry out a simulation study to compare the Bayesian methods based on full and conditional likelihood with standard frequentist approaches for case-crossover and time-series analysis. The proposed methods are illustrated through the Detroit Asthma Morbidity, Air Quality and Traffic study: a study to examine the association between acute asthma risk and ambient air pollutant concentrations.

e-mail: shili@umich.edu

COHERENT NONPARAMETRIC BAYES MODELS FOR NON-GAUSSIAN TIME SERIES

Zhiguang Xu, The Ohio State University
Steven MacEachern, The Ohio State University
Xinyi Xu*, The Ohio State University

Standard time series models rely heavily on assumptions of normality. However, many roughly stationary time series demonstrate a clear lack of normality of the limiting distribution. Classical statisticians are at a loss for how to deal with such data, resorting either to very restrictive families of transformations or passing to insuffient summaries of the data such as the ranks of the observations. Several classes of Bayesian nonparametric methods have been developed to provide flexible and accurate analysis for non-normal time series. Nevertheless, these models either do not incorporate serial dependence in series or cannot provide coherent refinement of the time scale. In this work, we propose a Bayesian copula model, which allows an arbitrary form for the marginal distributions while at the same time retain important properties of traditional time series models. This model normalizes the marginal distribution of the series through a cdf-inverse cdf transformation that relies on a nonparametric Bayesian component, and then places traditionally successful time series models on the transformed data. We demonstrate the properties of this copula model, and show that it provides better fit and improved short-range and long-range predictions than Gaussian competitors through both simulation studies and real data analysis.

e-mail: xinyi@stat.osu.edu

STATE-SPACE MODELS FOR COUNT TIME SERIES WITH EXCESS ZEROS

Ming Yang*, Harvard School of Public Health
Joseph Cavanaugh, University of Iowa
Gideon Zamba, University of Iowa

Time series comprised of counts are frequently encountered in many biomedical, epidemiological, and public health applications. For example, in disease surveillance, the occurrence of infection over time is often monitored by public health officials, and the resulting data are used for tracking changes in disease activity. For rare diseases with low infection rates, the observed counts typically contain a high frequency of zeros, reflecting zero-inflation. However, during an outbreak, the counts can also be very large, reflecting overdispersion. In modeling such data, failure to account for zero-inflation, overdispersion, and temporal correlation may result in misleading inferences and the detection of spurious associations. To facilitate the analysis of count time series with excess zeros, in the state-space framework, we develop a class of dynamic models based on either the zero-inflated Poisson (ZIP) or the zero-inflated negative binomial (ZINB) distribution. To estimate the model parameters, we devise a Monte Carlo Expectation Maximization (MCEM) algorithm, where particle filtering and particle smoothing methods are employed to approximate the high-dimensional integrals in the E-step of the algorithm. We consider an application based on public health surveillance for syphilis, a sexually transmitted disease (STD) that remains a major public health challenge in the United States.

e-mail: myang@sdac.harvard.edu

PENALIZED M-ESTIMATION AND AN ORACLE BLOCK BOOTSTRAP

Mihai C. Giurcanu*, University of Florida
Brett D. Presnell, University of Florida

Extending the oracle bootstrap procedure developed by Giurcanu and Presnell (2009) for sparse estimation of regression models, we develop an oracle block-bootstrap procedure for stationary time series. We show that the

oracle block-bootstrap and the m-out-of-n block-bootstrap estimators of the distributions of some penalized M-estimators are consistent and that the standard block-bootstrap estimators are inconsistent whenever the parameter vector is sparse. Using the parametric structure of the oracle block-bootstrap, we develop an efficient oracle block-bootstrap recycling algorithm that can be viewed as an alternative to the jackknife-after-bootstrap. An application to the estimation of the optimal block length, to the inference of a sparse autoregressive time series, and the results of a simulation study are also provided.

e-mail: giurcanu@stat.ufl.edu

EVOLUTIONARY FUNCTIONAL CONNECTIVITY IN fMRI DATA

Lucy F. Robinson*, Drexel University
Lauren Y. Atlas, New York University

Most existing techniques for functional connectivity networks inferred from functional fMRI data typically assume the network is static over time. We present a new method for detecting time-dependent structure in networks of brain regions. Functional connectivity networks are derived using partial coherence to describe the degree of connection between spatially disjoint locations in the brain. Our goal is to determine whether brain network topology remains stationary, or if subsets of the network exhibit changes over unknown time intervals. Changes in network structure may be related to shifts in neurological state, such as those associated with learning, drug uptake or experimental stimulus. We present a time-dependent stochastic blockmodel for fMRI data. Data from an experiment studying mediation of pain response by expectancy of relief are analyzed.

e-mail: lucy.f.robinson@gmail.com

A UNIFIED JOINT MODELING APPROACH FOR LONGITUDINAL STUDIES

Weiping Zhang, University of Science of Technology of China
Chenlei Leng, National University of Singapore
Cheng Yong Tang*, University of Colorado, Denver

In longitudinal studies, it is crucially important to synthetically understand the dynamics in the mean function, the variance function, and covariations of the repeated or clustered measurements. For the covariance structure, parsimoniously modeling approaches such as those utilizing the Cholesky type decompositions have been demonstrated effective in longitudinal data modeling and analysis. However, direct yet parsimonious approaches for revealing the covariance and correlation structures among longitudinal data remain less explored, and existing approaches may face difficulty when interpreting the correlation structures of longitudinal data. We propose in this paper a novel unified joint mean-variance-correlation modeling approach for longitudinal data analysis. By applying hyperspherical coordinates, we propose to model the correlation matrix of dependent longitudinal measurements by unconstrained parameters. The proposed modeling framework is parsimonious, interpretable, and flexible, and it guarantees the resulting correlation matrix to be non-negative definite. Extensive data examples and simulations support the effectiveness of the proposed approach.

e-mail: chengyong.tang@ucdenver.edu

118. HIERARCHICAL AND LATENT VARIABLE MODELS

BAYESIAN FAMILY FACTOR MODELS FOR ANALYZING MULTIPLE OUTCOMES IN FAMILIAL DATA

Qiaolin Chen*, University of California, Los Angeles
Robert E. Weiss, University of California, Los Angeles
Catherine A. Sugar, University of California, Los Angeles

The UCLA Neurocognitive Family Study collected more than 100 neurocognitive measurements on relatives of schizophrenia patients and relatives of matched control subjects, to study the transmission of vulnerability factors for schizophrenia. There are two types of correlations: measurements on individuals from the same family are correlated, and outcome measurements within subjects are also correlated. Standard analysis techniques for multiple outcomes do not take into account associations among members in a family, while standard analyses of familial data usually model outcomes separately and does not provide information about the relationship among outcomes. Therefore, I constructed new Bayesian Family Factor Models (BFFMs), which apply Bayesian inferences to confirmatory factor analysis (CFA) models with inter-correlated family-member factors and inter-correlated outcome factors. Results of model fitting on synthetic data show that the Bayesianfamily factor model can reasonably estimate parameters and that it works as well as the classic confirmatory factor analysis model using SPSS AMOS.

e-mail: qlchen@ucla.edu

MULTIVARIATE LONGITUDINAL DATA ANALYSIS WITH MIXED EFFECT HIDDEN MARKOV MODELS

Jesse D. Raffa*, University of Waterloo
Joel A. Dubin, University of Waterloo

The heterogeneity observed in a multivariate longitudinal response can often be attributed to underlying unobserved disease states. We propose modeling such disease states using a hidden Markov model (HMM) approach. We expand upon previous work which incorporated random effects into hidden Markov models (HMMs) for the analysis of univariate longitudinal data to the setting of a multivariate longitudinal response. Multivariate longitudinal data is modeled using separate but correlated random effects between longitudinal responses of mixed data types. We use a computationally efficient Bayesian approach using Markov chain Monte Carlo (MCMC). Simulations were performed to evaluate the properties of such models under a variety of realistic situations, and we apply our methodology to a bivariate longitudinal response data from a smoking cessation clinical trial.

e-mail: jdraffa@uwaterloo.ca

IMPROVED ASSESSMENT OF ORDINAL TRANSITIONAL DATA IN MULTIPLE SCLEROSIS THROUGH BAYESIAN HIERARCHICAL POISSON MODELS WITH A HIDDEN MARKOV STRUCTURE

Ariana Hedges*, Brigham Young University
Brian Healy, Massachusetts General Hospital
David Engler, Brigham Young University

Multiple sclerosis (MS) is one of the most common neurologic disorders in young adults in the United States, affecting approximately 1 out of every 1000 adults. One of the defining characteristics of MS is the heterogeneity in disease course across patients. The most common measure of disease severity and progression in MS is the expanded disability status scale (EDSS), which is a 0-10 ordinal scale in half-unit increments that combines information across seven functional systems. Markov transition models have been frequently employed to model transitions on the EDSS. However, because patient assessment on the EDSS scale can be error-prone with high interrater variability, it may be the case that a patient's true disease state is best modeled as an underlying latent variable under a hidden Markov model. We adopt a Poisson hidden Markov model (PHMM) under a Bayesian hierarchical framework to model EDSS transitions. Results are assessed both through simulation studies and through analysis of data collected from the Partners MS Center in Boston, MA.

e-mail: engler@byu.edu

IMPROVING THE ESTIMATE OF EFFECTIVENESS IN HIV PREVENTION TRIALS BY INCORPORATING THE EXPOSURE PROCESS

Jingyang Zhang*, Fred Hutchinson Cancer Research Center
Elizabeth R. Brown, Fred Hutchinson Cancer Research Center and University of Washington

Estimating the effectiveness of a new intervention is usually the main objective for HIV prevention trials. The Cox proportional hazard model is mainly used to estimate effectiveness, but it assumes that every subject has the same exposure process. Here we propose an estimate of effectiveness adjusted for the heterogeneity of exposure to HIV in the study population, using a latent Poisson process model for the exposure path of each subject. Moreover, we also consider the scenario in which a proportion of participants are not exposed to HIV by assuming a zero-inflated distribution for the rate of the exposure process. We employ a Bayesian estimation approach to estimate the exposure-adjusted effectiveness with informative prior for the per-act risk of infection under exposure, which we have adequate prior information. Simulation studies are carried out to validate the approach and explore the properties of the estimate. Sensitivity analyses are also performed to assess the proposed model.

e-mail: jzhang2@fhcrc.org

WEIGHTED KAPLAN-MEIER AND COMMENSURATE BAYESIAN MODELS FOR COMBINING CURRENT AND HISTORICAL SURVIVAL INFORMATION

Thomas A. Murray*, University of Minnesota
Brian P. Hobbs, University of Texas MD Anderson Cancer Center
Theodore Lystig, Medtronic Inc.
Bradley P. Carlin, University of Minnesota

Trial investigators often have a primary interest in the estimation of the survival curve in a population for which there exists acceptable historical information from which to borrow strength. However, borrowing strength from a historical trial that is systematically different from the current trial can result in biased conclusions and possibly longer trials. This paper develops an ad hoc method and a fully Bayesian method that attenuate bias and increase efficiency by automatically discerning the commensurability of the two sources of information using an extension of the Kaplan-Meier estimator and an extension of the commensurate prior approach for a flexible piecewise exponential proportional hazards model, respectively. The performance of these models regarding survival curve estimation is compared with other common strategies undertaken in the presence of acceptable historical information. We use simulation to show that these methods provide an attractive bias-variance tradeoff in a variety of settings. We also fit our methods to a pair of datasets from clinical trials on colon cancer treatment and assess the resulting improvement in estimate precision. We finish with a discussion of the advantages and limitations of these two methods for evidence synthesis, as well as directions for future work in this area.

e-mail: 8tmurray@gmail.com

119. COMPUTATIONAL METHODS AND IMPLEMENTATION

ORTHOGONAL FUNCTIONS IN THE STUDY OF VARIOUS BIOLOGICAL PROBLEMS

Mohsen Razzaghi*, Mississippi State University

The available sets of orthogonal functions can be divided into three classes. The first includes set of piecewise constant basis functions (e.g., Walsh, block-pulse, etc.). The second consists of set of orthogonal polynomials (e.g., Laguerre, Legendre, Chebyshev, etc.). The third is the widely used set of sine-cosine functions in Fourier series. While orthogonal polynomials and sine-cosine functions together form a class of continuous basis functions, piecewise constant basis functions have inherent discontinuities or jumps. In the mathematical modeling of biological processes, images often have properties that vary continuously in some regions and discontinuously in others. Thus, in order to properly approximate these spatially varying properties it is necessary to use approximating functions that can accurately model both continuous and discontinuous phenomena. For these situations hybrid functions, which are combinations of piecewise constant basis functions and continuous basis functions, will be more effective. In this work, we present a new approach to the solution of various biological problems. Our approach is based upon hybrid functions, which are combinations of block-pulse functions and Legendre polynomials. Numerical examples are included to demonstrate the applicability and the accuracy of the proposed method.

e-mail: razzaghi@math.msstate.edu

IMAGE DETAILS PRESERVING IMAGE DENOISING BY LOCAL CLUSTERING

Partha Sarathi Mukherjee*, Boise State University
Peihua Qiu, University of Minnesota

With rapid growth of the usage of images in many disciplines, preservation of the details of image objects while denoising becomes a hot research area. Images often contain noise due to imperfections of the image acquisition techniques. Noise in images should be removed so that the details of the image objects e.g., blood vessels or tumors in human brain are clearly seen, and the subsequent image analyses are reliable. Most image denoising techniques in the literature are based on certain assumptions about the continuity of the image intensities, edge curves (in 2-D images) or edge surfaces (in 3-D images) etc., which are often not reasonable in case the image resolution is low. If there are lots of details in the images including complicated edge structures, then these methods blur those. This talk presents a novel image denoising method which is based on local clustering. The challenging task of preserving image details including complicated edge structures is accomplished by performing local clustering and adaptive smoothing. The proposed method preserves most details of the image objects well even if the image resolution is not too high. Theoretical properties and numerical studies show that it works well in various applications.

e-mail: parthamukherjee@boisestate.edu

MAPPING QUANTITATIVE TRAIT LOCI UNDERLYING FUNCTION-VALUED PHENOTYPES

Il-Youp Kwak*, University of Wisconsin, Madison
Karl W. Broman, University of Wisconsin, Madison

Most statistical methods for QTL mapping focus on a single phenotype. However, multiple phenotypes are commonly measured, and recent technological advances have greatly simplified the automated acquisition of numerous phenotypes, including function-valued phenotypes, such as height measured over time. While there exist methods for QTL mapping with function-valued phenotypes, they are generally computationally intensive and focus on single-QTL models. We propose two simple fast methods that maintain high power and precision and are amenable to extensions with multiple-QTL models using the penalized likelihood approach of Broman and Speed (2002). After identifying multiple QTL by these approaches, we can view the function-valued QTL effects to provide a deeper understanding of the underlying processes. Our methods have been implemented as a package for R.

e-mail: ikwak2@wisc.edu

BIAS CORRECTION WHEN SELECTING THE MINIMAL-ERROR CLASSIFIER FROM MANY MACHINE LEARNING MODELS IN GENOMIC APPLICATIONS

Ying Ding*, University of Pittsburgh
Shaowu Tang, University of Pittsburgh
George Tseng, University of Pittsburgh

Supervised machine learning (a.k.a. classification analysis) is commonly encountered in genomic data analysis. When an independent testing data set is not available, cross validation is commonly used to evaluate an unbiased error rate estimate. It has been a common practice that many machine learning methods are applied to one data set and the method that produces the smallest cross-validation error rate is selected and reported. Theoretically such a minimal-error classifier selection produces bias with an optimistically smaller error, especially when the sample size is small and many classifiers are examined. In this paper, we applied an inverse power law (IPL) method to quantify the bias. Simulations and real applications showed a successful bias estimation through IPL. We further developed an application through a large-scale analysis of 390 data sets from GEO. We concluded a sequence of potentially high-performing classifiers a practitioner should examine and provide their associated bias range from empirical data given the sample size and number of classifiers tested. Finally, theory and statistical properties of the bias are demonstrated. Our suggested scheme is practical to guide future applications and provides insight to genomic machine learning problem.

e-mail: dingying85@gmail.com

CORRELATION BETWEEN TWO LARGE-SCALE TEMPORAL STATISTICS

Linlin Chen, Rochester Institute of Technology
Chen Ding*, University of Rochester

Modern computers all use cache memory. The performance may differ by 10 times depending on whether a program's active data fits in cache. It requires the monitoring of a large number of data accesses to characterize active data usage, because a program may make near a billion memory access each second (by a single core). Two window-based statistics have been used. The first is footprint, the amount of data used in a period of execution. For a program with N operations, the number of windows is quadratic (N choose 2). The second is reuse distance, which is the amount data accessed between two consecutive uses of the same datum. There are up to N reuse windows. For decades, the relation between them was not precisely established, partly because it was too costly to measure all footprint windows. Our recent work has developed new algorithms to efficiently measure the footprint in all quadratic number of windows. The initial results suggest a strong correlation between them. In this work, we use the industry standard benchmark suite to collect the base data, identify the cases of strong and weak correspondence, and statistically characterize the correlation or the lack of.

e-mail: cding@cs.rochester.edu

A NEW SEMIPARAMETRIC ESTIMATION METHOD FOR ACCELERATED HAZARDS MIXTURE CURE MODEL

Jiajia Zhang*, University of South Carolina
Yingwei Peng, Queen's University
Haifen Li, East China Normal University

The semiparametric accelerated hazards mixture cure model provides a useful alternative to analyze survival data with a cure fraction if covariates of interest have a gradual effect on the hazard of uncured patients. However, the application of the model may be hindered by the computational intractability of its estimation method due to non-smooth estimating equations involved. We propose a new semiparametric estimation method based on a smooth estimating equation for the model and demonstrate that the new method makes the parameter estimation more tractable without loss of efficiency. The proposed method is used to fit the model to a SEER breast cancer data set.

e-mail: jzhang@mailbox.sc.edu

THEORETICAL PROPERTIES OF THE WEIGHTED GENERALIZED RALEIGH AND RELATED DISTRIBUTIONS

Mavis Pararai*, Indiana University of Pennsylvania
Xueheng Shi, Georgia Institute of Technology
Broderick Oluyede, Georgia Southern University

A new class of weighted generalizations of the Raleigh distribution is constructed and studied. The statistical properties of these distributions including the behavior of the hazard or failure rate and reverse hazard functions, moments, moment generating function, mean, variance, coefficient of variation, coefficient of skewness and coefficient of kurtosis are obtained. Other important properties including entropy (Shannon and beta) which are measures of uncertainty in these distributions are also presented.

e-mail: pararaim@iup.edu

ENAR 2013
Spring Meeting
March 10 – 13

INDEX

12100 Sunset Hills Road | Suite 130
Reston, Virginia 20190
Phone 703-437-4377 | Fax 703-435-4390

ENAR