Abstracts

ENAR 2012 SPRING MEETING APRIL 1–4

WITH IMS AND SECTIONS OF ASA

Hyatt Regency Washington
on Capitol Hill – Washington, DC

1. BAYESIAN METHODS

1a. BAYESIAN MODELING OF ChIP-Seq DATA FOR DETECTING CHROMATIN REGIONS ATTACHED TO THE NUCLEAR ENVELOPE BASED ON LAMIN B1

Sabrina Herrmann, TU Dortmund University, Dortmund, Germany
Holger Schwender, TU Dortmund University, Dortmund, Germany*
Shoudan Liang, University of Texas, MD Anderson Cancer Center
Yue Lu, University of Texas, MD Anderson Cancer Center
Marcos Estecio, University of Texas, MD Anderson Cancer Center
Katja Ickstadt, TU Dortmund University, Dortmund, Germany Peter Mueller, University of Texas at Austin

The spatial organization of the chromosomes in the nucleus is influenced by chromatin regions binding to the nucleic lamina, i.e. the inner part of the nucleic envelope. To investigate the architecture of chromosomes in the interphase nucleus, it is thus of interest to detect such bound chromatin segments. This goal can be achieved by considering the fibrous protein Lamin B1 as a surrogate, since regions of high abundance of Lamin B1 can indicate chromatin segments attached to the nucleic lamina. To analyze data from a study in which Lamin B1 has been measured using the ChIP-Seq (Chromatin-Immunoprecipitation Sequencing) technology, we have developed a Bayesian procedure for modeling the change points in this data set. As we will show in our presentation, this approach is not only able to detect regions of high vs. low levels of Lamin B1, and therefore, gives an insight into the binding of the chromatin to the nucleic envelope, but also provides information on the variability in the segmentation. Moreover, this procedure is not restricted to Lamin B1 data, but can also be used for a binary segmentation in other (genetic) data and generalized to, for example, the analysis of copy number variations.

email: holger.schwender@udo.edu

1b. BAYESIAN LEARNING IN JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

Laura A. Hatfield, Harvard Medical School; James S. Hodges, University of Minnesota; Bradley P. Carlin, University of Minnesota*

In studying biological processes, investigators may repeatedly measure features of the process (longitudinal data) and also measure the time until some event (survival data). For example, a clinical trial may measure symptom severity and time until death. A broad class of joint models for simultaneously analyzing longitudinal and time-to-event data has been developed. The most popular such models use latent variables to link longitudinal and survival submodels. These approaches have expanded to accommodate many data complexities, yet little attention has been paid to these approaches' properties and performance. To quantify the benefit of joint versus separate modeling, we derive closed-form expressions for posterior quantities in a simplified normal-lognormal joint model. We show that as the prior variance on fixed effects increases, the resulting posterior means and variances from this joint model and either of its submodels alone converge to the same limit. We use a single latent variable to link the two submodels and thus require a coefficient to establish the magnitude and direction of its contribution to one of the submodels. The posterior for this scaling parameter may exhibit two modes symmetric about the origin when the sample size is small, indicating weak identification of its sign.

email: hatfield@hcp.med.harvard.edu

1c. BAYESIAN SEMIPARAMETRIC REGRESSION ANALYSIS OF BIVARIATE CURRENT STATUS DATA

Naichen Wang, University of South Carolina*
Lianming Wang, University of South Carolina

The Gamma-frailty proportional hazard model is widely used to analyze multivariate survival data in the literature. In this paper we first propose an easy-to-implement Bayesian method for analyzing bivariate current status data under the Gamma-frailty PH model. To avoid the restrictive parametric assumption for the frailty distribution, we extend the model such that it allows an unknown distribution for the frailty by adopting a Dirichlet process Gamma mixture prior for the distribution. An efficient Gibbs sampler is proposed based on the exact block Gibbs sampler. Our proposed method shows good performance in an extensive simulation study, and is applied to a tumorigenicity study conducted by National Toxicology Program.

email: cow1029@gmail.com

1d. A PHASE I TRIAL DESIGN FOR INCORPORATING EFFICACY OUTCOMES THAT ARE CONDITIONAL UPON ABSENCE OF DOSE-LIMITING TOXICITIES

Thomas M. Braun, University of Michigan; Shan Kang, University of Michigan*
Jeremy M G Taylor, University of Michigan

We propose a Phase I trial design in which there are three possible outcomes for each patient: dose-limiting toxicity (DLT), absence of therapeutic response without DLT, and presence of therapeutic response without DLT. We define the latter outcome as a "success." The goal of the trial is to find the most successful dose (MSD), the dose with the largest probability of success. We propose a design that accumulates information on patients with

regard to both DLT and response conditional on no DLT. Bayesian methods are used to update the estimates of DLT and response probabilities when each patient is enrolled, and use these methods to determine the dose level assigned to each patient. Due to the need to explore doses more fully, each patient is not necessarily assigned the current estimate of the MSD; our algorithm instead will assign a dose that is in a neighborhood of the current MSD. We examine the ability of our design to correctly identify the MSD in a variety of settings via simulation and compare the performance of our design to that of a competing approach.

email: shankang@umich.edu

1e. AN EMPIRICAL BAYES HIERARCHICAL MODEL FOR INFERENCE IN TIME-COURSE RNA-Seq EXPERIMENTS

Ning Leng, University of Wisconsin-Madison*
Victor Ruotti, Morgridge Institute for Research
Ron M. Stewart, Morgridge Institute for Research
James A. Thomson, Morgridge Institute for Research
Christina Kendziorski, University of Wisconsin-Madison

RNA-sequencing is a powerful approach providing estimates of both isoform and gene expression with unprecedented dynamic range and accuracy. A fundamental goal of RNA-seq experiments measuring expression in two or more biological conditions over time is the identification of the temporal patterns of isoform and gene level differential expression. Most of the statistical methods developed to identify differentially expressed genes over time measured using microarrays do not directly apply, and the methods that have been developed specifically for RNA-seq measurements do not accommodate sequencing bias, coding region information, dependence across isoforms, and dependence over time. We have developed an empirical Bayesian modeling approach that accounts for and capitalizes on these features. Advantages of the approach are illustrated in simulations and in an RNA-seq time-course study of human development done in collaboration with Jamie Thomson's lab at UW-Madison.

email: nleng@wisc.edu

1f. BAYESIAN INDIRECT AND MIXED TREATMENT COMPARISONS ACROSS LONGITUDINAL TIME POINTS

Ying Ding, Eli Lilly and Company*
Haoda Fu, Eli Lilly and Company

Meta-analysis has become an acceptable and powerful tool for pooling quantitative results from multiple studies addressing the same question. 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. The proposed model only uses summary level longitudinal data. This model is particularly useful when a meta-analysis is performed on studies with different durations. It enables the borrowing of information from shorter studies even in the situation where the primary interest is in a time point beyond the duration of some shorter 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: dingyi@lilly.com

1g. SAMPLE SIZE ESTIMATION FOR JOINT MODELING OF EFFICACY AND SAFETY

Brandi Falley, Baylor University*
James Stamey, Baylor University

Interest lies in simultaneously estimating efficacy and safety where the safety variable is subject to underreporting. We propose a Bayesian sample size determination method to account for the underreporting and appropriately power the study.

email: brandi_falley@baylor.edu

1h. IMPLEMENTATION OF CONTINUOUS BAYESIAN INTERIM MONITORING FOR SINGLE ARM PHASE II TRIALS WITH THE ONCORE SYSTEM

Stacey Slone, Markey Cancer Center, University of Kentucky*
Emily Van Meter, Markey Cancer Center, University of Kentucky
Dennie Jones, Markey Cancer Center, University of Kentucky

One of the latest advances in interim monitoring for single arm Phase II trial designs is Johnson & Cook's BFDesigner software. As presented in their Clinical Trials article from 2009, the software provides operating characteristics for a single arm Phase II trial with N patients for binary or right-censored time to event data under various assumptions. It also displays continuous stopping rules for futility and for superiority. While sufficient for protocol development, these types of continuously monitored trials require action whenever a new patient is enrolled or an event of interest occurs. Therefore, real-time notification from the data management program used at the institution is critical for interaction with the trial biostatistician and successful monitoring. We will present a SAS® program that will access the data from

the Oncore® data management system in real time and output the action needed for the trial along with the current operating characteristics. We provide an example of how this program could be utilized for a phase II lung cancer trial currently under development at the University of Kentucky Markey Cancer Center.

email: stacey.slone@uky.edu

1i. JOINT MODELING OF TIME-TO-EVENT AND TUMOR SIZE

Weichao Bao, University of South Carolina*
Bo Cai, University of South Carolina

In clinical trials, time-to-event data and longitudinal data are often collected. To model both the time-to-event and longitudinal components simultaneously, a joint modeling approach becomes increasingly important which can reduce potential biases and improve the efficiency in estimating treatment effects. In cancer clinical trials, change in tumor size is an important efficacy outcome which might serve as a surrogate for the overall survival time. In this paper, we propose a joint model where a nonlinear mixed-effect model is used to describe change in tumor size and an accelerated failure time model is used to describe overall survival time. The nonlinear mixed-effect model includes an exponential shrinkage and a linear progression. The survival and longitudinal components are linked through both fixed effects and random effects with appropriate adjustments. A simulation study is presented to evaluate the performance of the proposed approach compared with other competing models.

email: baow@email.sc.edu

1j. BAYESIAN ORDER RESTRICTED INFERENCE OF MEASUREMENT AGREEMENT WITH AN APPLICATION TO THE PHYSICIAN RELIABILITY STUDY

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

The Physician Reliability Study is an NICHD-sponsored trial that aims at estimating, and delineating factors of, intra- and inter-physician agreement in diagnosing the disease of endometriosis. As part of the study design, physicians reviewed clinical information of study participants in four settings, with each successive setting providing one more piece of clinical information than the one before it. An interesting question related to this design is whether the inter-physician agreement, as measured by the kappa statistic, is monotonically increasing with the setting (i.e., with increasing amount of clinical information). In this project, we consider a Bayesian test of this order restricted hypothesis and propose ways of estimating kappa statistics when this order restriction is assumed true. We also explore methods that allow certain degree of uncertainty in accounting for the constraint.

email: chenzhe@mail.nih.gov

1k. QUANTAL RESPONSES OF THE WEIBULL RISK FUNCTION

Douglas Moore, University of North Carolina at Wilmington*

Benchmark dose estimation is a widely used methodology in toxicology. It is one of the methods used by the EPA in determining allowable dose amounts of hazardous agents. This research focuses on quantal responses in the Weibull risk function in a Bayesian framework. However, to make the problem more applicable to researchers, we reparametrize using more intuitive parameters. In this talk, we will develop the methodology and illustrate the usefulness of this approach on several toxicological data sets.

email: dbm5951@uncw.edu

1l. BAYESIAN EFFECT ESTIMATION ACCOUNTING FOR ADJUSTMENT UNCERTAINTY

Chi Wang, University of Kentucky*
Giovanni Parmigiani, Dana-Farber Cancer Institute and Harvard School of Public Health
Francesca Dominici, Harvard School of Public Health

Model-based estimation of the effect of an exposure on an outcome is generally sensitive to the choice of which confounding factors are included in the model. We propose a new approach, which we call Bayesian Adjustment for Confounding (BAC), to estimate the effect on the outcome associated with an exposure of interest while accounting for the uncertainty in the confounding adjustment. Our approach is based on specifying two models: 1) the outcome as a function of the exposure and the potential confounders (the outcome model); and 2) the exposure as a function of the potential confounders (the exposure model). We consider Bayesian variable selection on both models and link the two by introducing a dependence parameter denoting the prior odds of including a predictor in the outcome model, given that the same predictor is in the exposure model. In the absence of dependence, BAC reduces to traditional Bayesian Model Averaging (BMA). In simulation studies we show that, in presence of dependence, BAC estimates the exposure effect with smaller bias than traditional BMA, and improved coverage. We then compare BAC, a recent approach of Crainiceanu et al. (2008), and traditional BMA in a time series data set of hospital admissions, air pollution levels and weather variables in Nassau, NY for the period 1999-2005. Using each approach, we estimate the short-term effects of PM2.5 on emergency admissions for cardiovascular diseases, accounting for confounding. This application illustrates the potentially significant pitfalls of misusing variable selection methods in the context of adjustment uncertainty.

email: chi.wang@uky.edu

1m. BAYESIAN SEMIPARAMETRIC REGRESSION MODELS FOR SEMICOMPETING RISKS DATA

Kyu Ha Lee, Harvard School of Public Health*
Sebastien Haneuse, Harvard School of Public Health
Francesca Dominici, Harvard School of Public Health

In clinical trials, patients are often exposed to two concurrent events; a non-terminal event and a terminal event. The special features of such data, referred to as semicompeting risks data, require special considerations in model development as the non-terminal event is censored by the terminal event but not vice versa. The two event times are generally expected to be correlated in many applications. Therefore, a survival model for semicompeting risks data is desired to account for dependency between bivariate failure times. For such purpose, several semiparametric models with dependence structure satisfying the gamma frailty copula have been developed under frequentist framework. However, only limited research has been done in the Bayesian paradigm. In this paper, we propose Bayesian semiparametric regression models for semicompeting risks data. We adopt the idea of shared frailty to incorporate a dependent structure between two failure times into our models. The cumulative hazard function is modeled a priori using a gamma process. One of the advantages in our approach is that the regression parameters and the dependence parameters are jointly modeled and estimated via Gibbs sampling. The performance of our Bayesian semicompeting risks models is evaluated in simulation studies. We apply our proposed methods to the problem of estimating re-hospitalization rates for pancreatic cancer patients.

email: klee@hsph.harvard.edu

1n. BAYESIAN RESTRICTED CONTOUR ESTIMATION METHOD FOR X-INACTIVATION RATIO FROM PYRO-SEQUENCING DATA

Alan B. Lenarcic, University of North Carolina at Chapel Hill*
John Calaway, University of North Carolina at Chapel Hill
Fernando de Pardo, University of North Carolina at Chapel Hill
William Valdar, University of North Carolina at Chapel Hill

Normal cells express only one active X chromosome. Cells belonging to females must deactivate one X per cell. Until 4th day of development each embryonic cell expresses both X's but then makes a semi-stochastic choice to silence one; all daughter cells replicate this inactivation. The average ratio of activation, X_mother/X_father in developed tissues deviates deterministically from 50/50, dependent on the cross of parent inbred strains. The hypothesis stands that this deviation is controlled by multi-allelic gene Xce located on X. We used pyro-sequencing to determine X-inactivation ratios for mice arising from an F1 cross of the Collaborative Cross founder strains, aiming to better locate Xce, infer the total number of alleles, and test for parent-of-origin effects. We develop a Bayesian statistical model that infers a total-body cell-expression estimate for each assayed individual using a hierarchical Beta model. Doing so allows us to estimate the effective number of embryonic cells that contributed to each tissue at the time of X-inactivation. Adapting our own restricted-space slice-sampling procedure to investigate the posterior of the hierarchical data-generation model, we demonstrate conditional Markov Chain Monte Carlo sampling that provides robust inference on a complex experimental dataset.

email: alenarc@med.unc.edu

1o. MULTIPLE IMPUTATION OF LATENT COUNTS FROM HEAPED SELF-REPORTED MEASUREMENTS OF DAILY CIGARETTE CONSUMPTION

Sandra D. Griffith, University of Pennsylvania*
Saul Shiffman, University of Pittsburgh; Daniel F. Heitjan, University of Pennsylvania

Measures of daily cigarette consumption, like many self-reported numerical data, exhibit a form of measurement error termed heaping. This occurs when quantities are reported with varying levels of precision, often in the form of round numbers. As heaping can introduce substantial bias to estimates, conclusions drawn from data subject to heaping are suspect. A doubly-coded data set with both a conventional retrospective recall measurement (time-line follow-back) and an instantaneous measurement not subject to heaping (ecological momentary assessment), allows us to model the heaping mechanism. We apply these results to a new data set of daily cigarette consumption where only heaped cigarette counts are available, and multiply impute latent true cigarette counts suitable for standard analysis. Using Bayesian methodology in the framework of data augmentation, we treat the true cigarette counts as latent unobserved data and draw observations from their conditional posterior distribution. To account for the longitudinal nature of the data, we induce a correlation structure in the imputed counts. Using Bayesian posterior predictive checks, we examine model fit with both graphical and quantitative diagnostics.

email: sgrif@upenn.edu

1p. BAYESIAN GRAPHICAL MODELS IN EPIGENETIC APPLICATIONS

Riten Mitra, University of Texas, MD Anderson Cancer Center*
Peter Mueller, University of Texas at Austin
Yuan Ji, University of Texas, MD Anderson Cancer Center

We develop a Bayesian Markov Random Field Model for Histone modifications (HMs). HMs are important post-translational features and are believed to co-regulate biological processes such as gene expression. The model is built hierarchically from a prior on unknown networks. An important feature is that we model directly the dependence between the binary activation status of HMs. This led us to construct a general dependence structure

through an autologistic density conditioned on the graph. A critical computational hurdle was the evaluation of a normalization constant, which we tackled through a Monte Carlo integration scheme. Through simulation studies, we show the validity of our methods and report findings from a next-generation sequencing dataset. Lastly, we extend the graphical prior structure to include multiple graphs. The extended model allows joint estimation when the graphs are dependent. We suggest that such an approach could be extremely useful in differentiating between networks under related and yet different biological conditions.

email: riten82@gmail.com

1q. BAYESIAN NONPARAMETRIC ESTIMATION OF FINITE POPULATION QUANTITIES IN ABSENCE OF DESIGN INFORMATION ON NONSAMPLED UNITS

Sahar Zangeneh, University of Michigan

In Probability proportional to size (PPS) sampling, the sizes for nonsampled units are not required for the usual Horvitz-Thompson or Hajek estimates, and this information is rarely included in public use data files. Previous studies have shown that incorporating information on the sizes of the nonsampled units through semiparameteric models can result in improved estimates of finite population quantities. When the design variables that govern the selection mechanism, are missing, the sample design becomes informative and predictions need to be adjusted for the effect of selection. We present a general framework using Bayesian nonparametric mixture modeling with Dirichlet process priors for imputing the nonsampled size variables required for model-based estimation, when such information is not available to the statistician analyzing the data. We then utilize these imputed sizes in a two-step model-based framework for inference on the finite population mean and median. Finally, we assess the performance of our method in estimating the mean and median tax revenues in Swedish municipalities.

email: saharzz@umich.edu

2. SURVIVAL ANALYSIS

2a. COMMONALITY ANALYSIS FOR SURVIVAL DATA, WITH AN APPLICATION TO DATA FROM BREAST CANCER PATIENTS WITH NEWLY DIAGNOSED BRAIN METASTASES

Binglin Yue, Moffitt Cancer Center*
Xianghua Luo, University of Minnesota
Haitao Chu, University of Minnesota
Paul Sperduto, University of Minnesota

Commonality analysis is widely used in educational and social science research. It partitions the variation of the outcome variable in a multiple regression model into corresponding parts that are explained by covariates, individually or jointly. However, commonality analysis is seldom used in medical research, especially in survival data analysis, where censored observations are present. In this paper, we propose the use of commonality analysis for right censored survival data. The coefficient of determination (R2), derived from a generalized gamma distribution based accelerated failure time model (AFT-GG), is used to quantify the variation explained by covariates for log-scaled time to event. As a comparison, we also partition the generalized coefficients of determination (R2g) of the AFT-GG models and that of the Cox's proportional hazards models. We apply the proposed methods to data from breast cancer patients with newly diagnosed brain metastases. The proportions of the R2 or R2g, as well as the associated bootstrap confidence intervals, explained by each covariate or the combinations of two or more covariates are reported. Patients' Karnofsky Performance Status (KPS) and tumor subtype defined by estrogen receptor (ER) and human epidermal growth factor receptor 2 (HER2) explained the most variation in the log survival time, independently.

email: ann.yue001@gmail.com

2b. AN AIPCW ESTIMATOR OF THE CUMULATIVE INCIDENCE FUNCTION UNDER MULTIPLE COMPETING CENSORING MECHANISMS

Brian Sharkey, Harvard University*
Michael Hughes, Harvard University
Judith Lok, Harvard University

Competing risks occur in a time-to-event study in which a patient can experience one of several types of events. Traditional methods for handling competing risks data presuppose one censoring process, which is assumed to be non-informative. In a controlled clinical trial, censoring can occur for several reasons: some non-informative, others informative. We propose an estimator of the cumulative incidence function in the presence of multiple types of censoring mechanisms. The relationship between each

censoring process and the recorded prognostic variables can differ, especially if some censoring processes are informative (e.g. treatment non-adherence) and others are not (administrative censoring). Consequently, we incorporate this information by modeling each censoring process separately. We rely on semi-parametric theory to derive an augmented inverse probability of censoring weighted (AIPCW) estimator and its standard errors. We demonstrate the efficiency gained when using the AIPCW estimator compared to a non-augmented estimator via simulations. We then apply our method to evaluate the safety and efficacy of three HAART regimens in a study conducted by the AIDS Clinical Trial Group Network, ACTG A5142.

email: bsharkey@hsph.harvard.edu

2c. INCORPORATING SAMPLING PLAN AND COMPETING RISKS IN ANALYSIS OF PROSPECTIVE PREGNANCY STUDIES

Kirsten J. Lum, Johns Hopkins Bloomberg School of 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 with preconception enrollment are a valuable approach to quantifying exposure effects at critical windows in human reproduction and development. Couples are enrolled prior to or shortly after discontinuing contraception, thereby observing menstrual cycles at risk for pregnancy. A question that arises in these and other, possibly-prevalent cohort studies is what should be done with menstrual cycles that are not completely observed. Such cycles arise because enrollment occurred beyond the start of the at-risk cycle (left truncation). In addition, these cycles may be subject to censoring at a positive pregnancy test, or other censoring mid-cycle due to loss to follow-up. Here, we address enrollment cycles and pregnancy cycles and their contribution to the likelihood for estimating mean menstrual cycle length. We evaluate dropping the enrollment cycle for all women, conceding a loss of information versus modeling the full dataset, accounting for length selection bias. Furthermore, we model the joint distribution of cycle length and time to positive pregnancy test to account for censoring due to pregnancy. We evaluate these approaches in simulations informed by prospective pregnancy studies and compare estimates of mean menstrual cycle length in an analysis of the Longitudinal Investigation of Fertility and the Environment Study.

email: klum@jhsph.edu

2d. ESTIMATION OF COX PROPORTIONAL HAZARDS MODELS FOR TWO NEGATIVELY CORRELATED PROCESSES

Wenjing Xu, George Washington University*
Qing Pan, George Washington University
Joseph L. Gastwirth, George Washington University

In addition to its use in Biostatistics, this talk will demonstrate how survival analysis can model processes arising in equal employment litigations. In the context of discrimination in promotion, the lost opportunity may affect the individual's decision to retire, if eligible. This situation can be modeled by two Cox PH models, one for promotion conditional on not retiring and one for retiring. To account for the negative relationship between the two outcomes, besides the fixed covariates, a frailty term which increases the risk of one process (promotion) multiplicatively but decreases the risk of the other outcome (retirement) is introduced. Several frailty distributions are explored. An MCEM algorithm for fitting the joint process is proposed. The asymptotic properties of the parameter estimates are derived and verified through simulation studies. The sensitivity of the results to the choice of frailty distribution is examined. The advantages of modeling the negative relationship over fitting two marginal Cox PH models are demonstrated. The method is applied to data from an actual legal case.

email: jeanvane@gwmail.gwu.edu

2e. RECURSIVE PARTITIONING BASED WEIGHTS FOR CENSORED QUANTILE REGRESSION

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

Cox proportional hazards regression is the most commonly used survival analysis method for adjusted analysis in practice today. Yet, the difficult interpretation of the hazard ratio, due partly to its relative measure of association, motivates the use of a direct summary measure with interpretation based on units of time, such as change in median survival time. Censored quantile regression is emerging as a viable alternative/complement to the Cox model by providing a direct interpretation in terms of difference in quantile survival times. Current censored quantile methods rely upon the fairly strong assumptions of unconditionally independent censoring or linearity in all quantiles. Recently, Wang and Wang (2009) proposed a locally weighted censored quantile regression approach utilizing kernel estimation that avoided assuming unconditionally independent censoring and linearity in all quantiles. We propose a similar approach that uses recursive partitioning to define locality, which we believe will be better able to handle high dimensional data analysis. We analyzed the performance of the proposed tree based approach via an extensive simulation study and found that tree based weights have potential to provide robust estimation at little loss of precision. We illustrate the proposed estimator using data from a clinical trial on primary biliary cirrhosis.

email: weyxx003@umn.edu

2f. DEVELOPMENT AND EVALUATION OF MULTIMARKER PANELS FOR CLINICAL PROGNOSIS

Benjamin French, University of Pennsylvania*
Paramita Saha Chaudhuri, Duke University
Bonnie Ky, University of Pennsylvania
Thomas P. Cappola, University of Pennsylvania
Patrick J. Heagerty, University of Washington

Clinical research regarding heart failure suggests that biomarkers could be combined with relevant clinical information to more accurately quantify risk heterogeneity among all patients and to inform more beneficial treatment strategies for individual patients. Therefore, statistical methodology is required to determine which multimarker panel provides improved prognostic metrics. We focus on analytic strategies to develop an optimal marker combination and to evaluate its accuracy in predicting a censored survival outcome. A Cox regression model can be used to develop a composite marker as a weighted combination of component markers, in which weights are determined by estimated regression coefficients. To evaluate predictive accuracy, the composite marker can be supplied as the input to a time-dependent ROC analysis. Alternatively, risk reclassification can be used to contrast the predictive accuracy of models with and without marker(s) of interest. We demonstrate these approaches using data from the Penn Heart Failure Study. In simulation studies, we evaluate the impact of a sub-optimal marker combination on estimation of the area under the ROC curve and the net reclassification improvement. We recommend that analysts examine the functional form for component markers and consider plausible forms for effect modification to determine whether the composite marker is optimally specified.

email: bcfrench@upenn.edu

2g. GENERALIZED ODDS-RATE HAZARD MODELS FOR INTERVAL-CENSORED FAILURE TIME DATA

Bin Zhang, University of Alabama-Birmingham*
Lianming Wang, University of South Carolina

Interval-censored data naturally occur in many fields and the feature is that the failure time of interest is not observed exactly but known to fall within some interval (Kalbfieisch and Prentice 2002; Sun 2006). In this paper, we propose a novel semiparametric generalized odds-rate hazard (GORH) models for analyzing interval-censored data as an alternative to existing semiparametric models in the literature. We propose to approximate the unknown nonparametric nondecreasing function in the GORH model with a linear combination of monotone splines, leading to only finite unknown parameters to estimate. The regression parameters and the baseline survival function are estimated jointly. The proposed methods work well as shown in the simulation study and are easy to implement. Two real-life interval-censored data are analyzed for illustration.

email: binzhang@uab.edu

2h. A SEMI-PARAMETRIC JOINT MODEL FOR SEMI-COMPETING RISK DATA

Renke Zhou, University of Texas, MD Anderson Cancer Center*
Jing Ning, University of Texas, MD Anderson Cancer Center
Melissa Bondy, Baylor College of Medicine

We propose an approach that uses a joint bivariate survival model for the semi-competing risk problem, in which a terminal event (usually death) censors an intermediate event (disease process landmark), but not vice versa. In the study about the distribution of the intermediate event, we cannot treat the terminal event as independent censoring since the two events are correlated and most likely there is some effect of the intermediate event on the residual survival. To investigate the pattern of association between those two events, a plot of conditional hazard ratio is used based on a nonparametric method in the upper wedge data where both of the events are observed. Then the proper Archimedean copula is selected to formulate the joint survival distribution. The frailty can be estimated by the pseudo-maximum likelihood of the two-stage semi-parametric method. Simulations are performed to check the model. We apply the model to the analysis of time to recurrence and time to death in the data of the Early Stage Breast Cancer Repository (ESBCR) cohort study from The University of Texas M.D. Anderson Cancer Center.

email: renke.zhou@uth.tmc.edu

2i. INFORMATIVE AGE REDUCTION MODEL FOR RECURRENT EVENT

Li Li, University of South Carolina*
Timothy Hanson, University of South Carolina

Repairable systems have been widely studied in literature. In reliability setting, a system fails and upon each failure time the system gets repaired; in clinical study, certain symptom shows up and a patient receives one of the possible treatments each time. Each repair (treatment) brings the system (patient) to a certain state of the life distribution by modeling age reduction (Kijima type I and II model) (Kijima, M. (1989)). Common assumption is that repairs in a category share the same proportion of age reduction, e.g. good as new, bad as old, or in between. However, this assumption is questionable due to the heterogeneity of each repair even within a category, e.g. the different components being replaced, quality of the new components, skills of the technicians, etc. Subsequently, we consider the distribution for the proportion of age reduction being linked to the potential covariates and further random effects. By quantifying the covariates effects on the repair (treatment), we are able to predict the actual effect of a future repair (treatment) and further quantify the conditional reliability (survival) distribution for the system.

email: lil@email.sc.edu

2j. PARAMETER ESTIMATION IN COX PROPORTIONAL HAZARD MODELS WITH MISSING CENSORING INDICATORS

Naomi C. Brownstein, University of North Carolina at Chapel Hill*
Eric Bair, University of North Carolina at Chapel Hill
Jianwen Cai, University of North Carolina at Chapel Hill
Gary Slade, University of North Carolina at Chapel Hill

In a prospective cohort study, examining all participants for incidence of the condition may be prohibitively expensive. For example, the "gold standard" for diagnosing temporomandibular disorder (TMD) is a clinical examination by an expert dentist. In a large study, examining all subjects in this manner is infeasible. Instead, it is common to use a cheaper (and less reliable) examination to screen for possible incident cases and perform the "gold standard" examination only on participants who screen positive on this cheaper examination. Unfortunately, some subjects leave the study before receiving the "gold standard" examination. In the context of survival analysis, this results in missing censoring indicators. We propose a method for parameter estimation in survival models with missing censoring indicators. We impute the probability of being a case for those with no "gold standard" examination using the EM algorithm. These estimated probabilities are used to estimate the hazard ratios associated with each putative risk factor. The variance introduced by the imputation is estimated using bootstrapping. We simulate data with missing censoring indicators and show that our method performs better than the competing methods. We also apply our proposed method to a large prospective cohort study of TMD.

email: nbrownst@email.unc.edu

2k. CHALLENGES FROM COMPETING RISKS AND RECURRENT EVENTS IN CARDIOVASCULAR DEVICE TRIALS: A REGULATORY REVIEWER'S PERSPECTIVE

Yu Zhao, Center for Devices and Radiological Health, U.S. Food and Drug Administration*

All-cause-mortality and disease-free-survival are commonly used survival endpoints in clinical trials. However, in many cardiovascular device trials, multiple types of clinical events are often of interest and different types of events are of different clinical importance. Moreover, nonfatal events of interest, such as hospitalization, may occur repeatedly during the follow-up. Therefore, the aforementioned two approaches might not be of the most interest. In this presentation, I will present and discuss several approaches utilized in the recent cardiovascular device premarket submissions when facing the challenges from competing risks and recurrent events.

email: yu.zhao@fda.hhs.gov

2l. BAYESIAN SEMIPARAMETRIC MODEL FOR SPATIAL INTERVAL-CENSORED FAILURE TIME 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 data are often collected in practice. Although some methods are developed for analyzing such data, some issues still remain in terms of efficiency and accuracy of the estimations. In addition, interval-censored data with spatial correlation are not unusual but less studied. In this paper, we propose an efficient Bayesian method under proportional hazards model to analyze interval-censored data with spatial correlation. Specifically, a linear combination of monotone splines is used to model the baseline cumulative hazard function. This method specifies a finite number of parameters while still allows for great modeling flexibility. Data augmentation through Poisson latent variables is used to facilitate the derivation of posterior distributions that are essential in the Gibbs sampler proposed. A conditional autoregressive (CAR) prior is employed to model the spatial frailties. A simulation study is conducted to evaluate the behavior of the proposed method. Smoking cessation data where the subjects reside in 54 zip code areas in southeast Minnesota are analyzed as an illustration.

email: Chunpan2003@hotmail.com

2m. ANALYSIS OF VARIANCE FOR RIGHT CENSORED SURVIVAL DATA

Chetachi A. Emeremni, University of Pittsburgh*

Analysis of variance has been one of the most powerful statistical tools for comparing mean continuous response across multiple groups. Use of classical ANOVA in time-to-event data is problematic because of the right censored nature of survival times. In this paper, we propose a weighted analysis of variance approach to comparing mean continuous response between groups when the outcome is subject to right censoring. The method weights each observation by the inverse of the probability of being censored. We show that classical ANOVA methods such as decomposition of sums of squares and tests of contrasts follows in the weighted ANOVA setting. Simulation results show that the weighted ANOVA could be a comparable alternative to other methods of analyzing survival data. We apply our methods to a dataset from the Radiation Therapy Oncology Group.

email: che7@pitt.edu

2n. SAMPLE SIZE AND POWER CALCULATION FOR PROPORTIONAL HAZARDS MODEL WITH TIME-DEPENDENT VARIABLES

Songfeng Wang, University of South Carolina*
Jiajia Zhang, University of South Carolina
Wenbin Lu, North Carolina State University

Cox proportional hazards (Cox PH) model with time-independent variables (sometimes referred as extended PH model) has been widely used in medical and clinical studies. Theories and practices regarding model estimation and fitting have been well developed for extended model. However, little work has been done in the design aspects. Sample size and power calculation is a very important topic in designing randomized clinical trials. In this paper, we develop a sample size formula based on the PH model with time-dependent variables by investigating the asymptotic distributions of the standard weighted log-rank statistics under the null and local alternative hypotheses. The derived sample size formula is an extension of Schoenfeld's sample size formula for the standard Cox PH model. Furthermore, the impacts of accrual methods and durations of accrual and follow-up periods on sample size are also investigated as numerical examples. The performance of the proposed formula is evaluated by extensive simulation studies and examples are given to illustrate its application using real data.

email: songfeng@gmail.com

3. STATISTICAL GENETICS/GENOMICS

3a. SIMULTANEOUS FUNCTIONAL CATEGORY ANALYSIS

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

An important task in statistical genomics is to integrate exogenous functional information with experimental genomic data, which provides insights to explain the data with a summary of functional content. Most category-analysis methods are designed to consider one category at a time, leaving the task of prioritizing categories as a secondary problem. There are definite advantages to methods that consider all categories simultaneously, since overlaps and size variation can be accommodated, but available methods are especially challenging computationally. In this work we develop an approach based on regression modeling. We investigate the close relationship between this regression approach and available model-based methods. The advantages of this method are illustrated in simulation studies and the application to identify functional categories that are important to Influenza virus replication.

email: he@stat.wisc.edu

3b. COMBINING LINKAGE ANALYSIS AND NEXT GENERATION SEQUENCING TO IDENTIFY RARE CAUSAL VARIANTS IN COMPLEX DISEASES

Silke Szymczak, National Human Genome Research Institute, National Institutes of Health*
Qing Li, National Human Genome Research Institute, National Institutes of Health
Claire L. Simpson, National Human Genome Research Institute, National Institutes of Health
Robert Wojciechowski, Johns Hopkins Bloomberg School of Public Health
Xilin Zhao, National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health
MaryPat S. Jones, National Human Genome Research Institute, National Institutes of Health
Richa Agarwala, National Center for Biotechnology Information, National Institutes of Health
Alejandro A. Schaeffer, National Center for Biotechnology Information, National Institutes of Health
Stephen A. Wank, National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health
Joan E. Bailey-Wilson, National Human Genome Research Institute, National Institutes of Health

Investigation of the common disease-rare variant hypothesis is now feasible using next generation DNA sequencing technologies. However, sequencing all exons or whole genomes of many individuals usually identifies thousands of rare variants. One promising approach to distinguish causal variants from sequencing artifacts or variants without functional effects is to focus on regions implicated by classic linkage analysis. Simulated sequence data in families is available from the Genetic Analysis Workshop 17. We show that power to detect causal variants substantially increases if linkage analysis is based on extended pedigrees instead of sib pairs or nuclear families. We illustrate our approach with linkage and sequencing data from a tumor syndrome study.

email: silke.szymczak@nih.gov

3c. USING GROWTH MIXTURE MODELING TO IDENTIFY LOCI ASSOCIATED WITH THE PROGRESSION OF DISEASE

Tong Shen, Duke University*

In a genome-wide association study (GWAS) for a longitudinal quantitative trait, the trait is measured at multiple time points. GWAS is the examination of marker loci to identify loci associated with the progression of the quantitative trait. I use a multi locus model, to simulate a longitudinal quantitative trait. I use the growth mixture modeling (GMM) method to assign each member of a sample into one of a small number of trajectory groups. The clinically important trajectory group is the one with fastest progression. The Bayesian posterior probability (BPP) of being in the clinically important group is used as a quantitative trait. I test for association with marker loci. I also use the modal BPP in the association test and perform a case/control association analysis. Finally, I compare these methods with the contingency

table method. I evaluate the empirical type I error and empirical power using null simulations and power simulations. The principal results are that: (1) Both the BPP method and modal BPP method maintain the correct type I error rate. (2) Both the BPP and modal BPP methods have significant power to detect the disease loci in the multi locus model. The powers of detecting a specific locus are proportional to minor allele frequency (MAF) of loci. (3) Powers drop at the loci which are in high LD with other loci.

email: tshen.stonybrook@gmail.com

3d. DATA PREPROCESSING: QUANTIFICATION AND NORMALIZATION OF THE LUMINEX ASSAY SYSTEM

Eileen Liao, University of California at Los Angeles*
David Elashoff, University of California at Los Angeles

In microarray experiments, variations in expression measurements emerge from many sources. We applied normalization methods to a bead-based multiplex Luminex assay system to reduce the plate-to-plate variation. Normalization on the Luminex assay system is a fundamental different scenario than the traditional affymetrix microarray. In affymetrix microarray the data we observe from each experimental unit is a vector as one gene expression corresponds to one subject. While in the Luminex system, each experimental unit is a plate and each plate has multiple subjects and analytes. Normalization across Luminex plates is normalization across matrices rather than a vector. We first quantified performance among measurements of fluorescent intensity, background in fluorescent intensity, and observed concentration in both high and standard scanning systems, and evaluated the preservation of ranking separation across plates. We then applied scale normalization, quantile normalization, lowess curve normalization to the Luminex system, and use the coefficient of variation across plates to measure each performance. We validate these methods by applying them to a separate lung transplant study.

email: biochen@gmail.com

3e. BORROWING INFORMATION ACROSS GENES AND EXPERIMENTS FOR IMPROVED RESIDUAL VARIANCE ESTIMATION IN MICROARRAY DATA ANALYSIS

Tieming Ji, Iowa State University*
Peng Liu, Iowa State University
Dan Nettleton, Iowa State University

Statistical inference for microarray experiments usually involves the estimation of residual variance for each gene. Because the sample size available for each gene is often low, the usual unbiased estimator of the residual variance can be unreliable. Shrinkage methods, including empirical Bayes approaches that borrow information across genes to produce more stable estimates, have been developed in recent years. Because the same microarray platform is often used for at least several experiments to study similar biological systems, there is an opportunity to improve variance estimation further by borrowing information not only across genes but also across experiments. We propose a lognormal model for residual variances that involves random gene effects and random experiment effects. Based on the model, we develop an empirical Bayes estimator of the residual variance for each combination of gene and experiment and call this estimator BAGE because information is Borrowed Across Genes and Experiments. A permutation strategy is used to make inference about the differential expression status of each gene. Simulation studies with data generated from different probability models and real microarray data show that our method outperforms existing approaches.

email: tji@iastate.edu

3f. JOINT MODELING OF DISEASE AND ENDOPHENOTYPE TO CHARACTERIZE THE EFFECT OF GENES AND THEIR INTERACTIONS

Alexandre Bureau, Université Laval - Robert-Giffard, Université Laval, Québec, Canada*
Jordie Croteau, Université Laval - Robert-Giffard, Université Laval, Québec, Canada
Molière Nguilé Makao, Université Laval - Robert-Giffard, Université Laval, Québec, Canada

To overcome phenotypic complexity in genetic studies of psychiatric disorders, a widespread approach is to measure endophenotypes, traits related to a disease and believed to be influenced by fewer genes. The presence of impairments on an endophenotype in both affected and non-affected relatives can provide information on familial transmission of alleles of genes involved in epistatic interactions causing the disease. This project aims at fulfilling the need for methods exploiting this information by jointly modeling a disease and its endophenotypes in relation to genes and their interactions. The disease and the endophenotype are modelled as dichotomous traits. We explore scenarios of genetic influences on the disease and endophenotype and propose

models to detect the genetic effects. We present and compare within-family conditional analysis and population level analysis of association between the genotypes of markers in two genes and the phenotype formed by the combination of the disease and endophenotype presence/absence. We derived score tests under the two approaches and compare their power under the various scenarios considered using simulations. The method is applied to cognitive endophenotypes and the schizophrenia and bipolar disorder diagnoses in extended families from Eastern Québec.

email: alexandre.bureau@msp.ulaval.ca

3g. EPISTASIS ENRICHED NETWORK AND RISK SCORE MODELING OF CONTINUOUS MULTIFACTOR DIMENSIONALITY REDUCTION

Hongying Dai, Children's Mercy Hospital*
Richard Charnigo, University of Kentucky
Mara Becker, University of Kentucky
Steve Leeder, University of Kentucky

Multifactor Dimensionality Reduction (MDR) has been widely applied to detect gene by gene (GxG) interactions associated with complex diseases. Existing MDR methods summarize disease risk by a dichotomous predisposing variable, which may limit accuracy in predicting the risk of disease. We propose a Continuous Multifactor Dimensionality Reduction (C-MDR) method that exhaustively searches for and detects significant GxG interactions to generate an epistasis enriched gene network. A continuous epistasis enriched risk score, which takes into account multiple GxG interactions simultaneously, replaces the dichotomous predisposing variable and provides higher resolution in the quantification of disease susceptibility. Application of the C-MDR method to a data set derived from Juvenile Idiopathic Arthritis patients treated with methotrexate (MTX) revealed several GxG interactions in the folate pathway that were associated with treatment response. The epistasis enriched risk score that pooled information from 82 significant GxG interactions distinguished MTX responders from non-responders with 82% accuracy. The proposed C-MDR is thus better able to distinguish between affected and unaffected subjects, especially in small samples. Simulation studies show that new GxG interaction measures (pOR, pRR and pChi) generate stronger power as compared to the existing MDR methods.

email: hdai@cmh.edu

3h. JOINT ANALYSIS OF SNP AND GENE EXPRESSION DATA IN GENOME-WIDE ASSOCIATION STUDIES

Yen-Tsung Huang, Harvard University*
Xihong Lin, Harvard University
Tyler VanderWeele, Harvard University

Genome-wide association studies (GWAS) have been a common practice in assessing the association between single nucleotide polymorphisms (SNPs) and disease phenotype. However, the causal mechanism between SNPs and disease is usually neglected. Here we propose to integrate the information of gene expression to reveal the mechanistic pathway between SNPs and disease. The relations among SNPs, gene expression and disease are modeled in the framework of causal mediation modeling. Using counterfactual approach, the direct and indirect effects of SNPs on disease can be derived. Furthermore, we propose a variance component test to investigate whether there exists an overall effect of SNPs on disease by borrowing gene expression information. The test statistic under the null follows a mixture of Chi-square distributions, which can be approximated with the scaled Chi-square distribution or with perturbation procedure. The relative performance of the tests for different disease models depends on the underlying true model. To accommodate different scenarios, we also construct an omnibus test. In simulation studies, our proposed test performs well and the omnibus test can almost reach the optimal power, in which the disease model is correctly specified. We apply our method to reanalyze the overall effect for the SNP set at ORMDL3 gene on the risk of asthma using MRC-A data (Moffatt et al., 2007).

email: ythuang@hsph.harvard.edu

3i. A ROBUST TEST FOR DETECTING DIFFERENTIALLY METHYLATED REGIONS

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

A major result of recent genome-wide association studies is that common variations in the primary DNA sequences explains a small portion of the genetic predisposition to common complex diseases. Epigenetic changes hold promise of explaining a portion of the missing heritability. Advances in next-generation sequencing have made epigenome profiling feasible. However, it poses substantial challenges to statistical analysis. Most of current statistical methods developed for detecting differential expression with microarrays are not directly applicable because of the unknown distribution of methylation status. Moreover, they could only detect differential methylation at each CpG site. Yet biologically it is more relevant to detect differentially methylated patterns in a region such as a CpG island or shore region. It is statistically challenging

because of the correlation of methylation status of CpG sites in a region and the unknown distribution of methylation and its pattern across sites in a region. We develop a robust test for detecting differentially methylated regions based on contrasting the squared difference of methylation proportions between cases and controls. Simulation results and application to real data sets from an epigenome study of obesity show that it can effectively detect differentially methylated regions.

email: hxu@georgiahealth.edu

3j. TESTING FOR GENE-ENVIRONMENT AND GENE-GENE INTERACTIONS UNDER MONOTONICITY CONSTRAINTS

Summer S. Han, National Cancer Institute, National Institutes of Health*
Philip S. Rosenberg, National Cancer Institute, National Institutes of Health
Nilanjan Chatterjee, National Cancer Institute, National Institutes of Health

Recent genome-wide association studies (GWAS) for the detection of main effects of genetic markers and their replications have had considerable success. However, relatively few gene-gene or gene-environment interaction findings have been successfully reproduced. Besides the main issues regarding insufficient sample size in current studies, a complication is that interactions that rank high based on p-values often correspond to extreme forms of joint effects that are biologically less plausible. To reduce false positives and to increase power, we have developed various gene-environment/gene-gene tests based on biologically more plausible constraints using bivariate isotonic regressions for case-control data. We extend our method to exploit gene-environment or gene-gene independence information, integrating the approach proposed by Chatterjee and Carroll (2005). We propose appropriate non-parametric and parametric permutation procedures for evaluating significance of the tests. Simulations show that our method gains power over traditional methods by reducing the sizes of alternative parameter spaces. We apply our method to several real data examples, including an analysis of bladder cancer data to detect interactions between the NAT2 gene and smoking.

email: summer.han@aya.yale.edu

3k. BAYESIAN GENE SET TEST, THE PROPORTION OF SIGNIFICANT GENES IN THE SET AS THE SUMMARY STATISTIC

Di Wu, Harvard University*
Ke Deng, Harvard University
Ming Hu, Harvard University
Jun Liu, Harvard University

Gene set tests are used in gene expression data analysis. They have better power than single gene tests, and help to understand the biological pathways and relate two datasets. One of the critical steps is to choose a suitable summary set statistic. After that, the significant level of the statistic is obtained by permutation or parametric approximation. The available gene set tests usually use the mean statistics (mean, mean of square, max-mean etc.) as the summary statistic. They also care the relationship between the summary statistics and the proportion of significant genes in the set. The proportion is an interesting feature of a gene set; however, no methods actually use it as a summary statistic due to the difficulty to capture the distribution of proportion. The significance of proportion also depends on the size of the test. The problems of p-values, caused by gene permutation (anti-conservative) and sample permutation (limited by sample size), remain here. We propose a novel Bayesian method and use MCMC to solve these problems for the proportion summary statistics.

email: dwu@fas.harvard.edu

3l. ADJUSTMENT FOR POPULATION STRATIFICATION VIA PRINCIPAL COMPONENTS IN ASSOCIATION ANALYSIS OF RARE VARIANTS

Yiwei Zhang, University of Minnesota*
Weihua Guan, University of Minnesota
Wei Pan, University of Minnesota

Principal component (PC) analysis has been shown to be able to adjust for population stratification in association analysis of common variants (CVs), while it is less clear how it would perform in association analysis of rare variants (RVs). Furthermore, with next-generation sequencing data, it is unknown whether PCs should be constructed based on CVs or RVs. In this study, we use the 1000 Genome Project sequence data to explore the construction of PCs and their use in association analysis of RVs. In particular, we demonstrate that PCs based on RVs is effective in controlling Type I error rates when applied to several association tests.

email: zhan1447@umn.edu

3m. REPRIORITIZING GENETIC ASSOCIATIONS IN HIT REGIONS USING LASSO-BASED RESAMPLE MODEL AVERAGING

William Valdar, University of North Carolina at Chapel Hill
Jeremy Sabourin, University of North Carolina at Chapel Hill*
Andrew Nobel, University of North Carolina at Chapel Hill
Chris Holmes, University of Oxford, United Kingdom

Significance testing using single predictor logistic models has proven useful for identifying genomic regions that harbor variants affecting human disease. But once a 'hit region' of association has been identified, local correlation due to linkage disequilibrium (LD) can make the set of underlying true signals ambiguous. Simultaneous modeling of multiple loci should help, but is seldom applied in a principled fashion. When it is, it typically includes no assessment of how sensitive model choice was to sampling variability. We present a general method for characterizing uncertainty in model choice that is well suited to reprioritizing SNPs within a hit region under strong LD. Our method, LLARRMA, combines LASSO shrinkage with resample model averaging and multiple imputations, estimating for each SNP the probability that it would be included in a multi-SNP model in alternative realizations of the data. We apply LLARRMA to simulations based on case-control GWAS data, and find that when there are several causal SNPs and strong LD, LLARRMA identifies a set of candidates that is enriched for causal loci relative to single locus analysis and the recently proposed method of Stability Selection.

email: jsabouri@unc.edu

3n. MULTILAYER CORRELATION STRUCTURE OF MICROARRAY GENE EXPRESSION DATA

Linlin Chen, Rochester Institute of Technology*
Lev Klebnov, Charles University
Anthony Almudevar, University of Rochester

In this project, we focus on possible causes of between-gene dependencies and their effects on the performance of gene selection procedures. We show that there are at least two noise-type reasons for high correlations between gene expression levels. First is of technical character, and is connected to a random character of the number of cells used to prepare microarray. Another reason is the heterogeneity of cells in a tissue. Both reasons allow one to make some predictions, which are verified on real data.

email: linlin.chen@gmail.com

3o. STATISTICAL METHODS FOR IDENTIFYING BATCH EFFECTS IN COPY NUMBER DATA

Sarah E. Reese, Virginia Commonwealth University*
Terry M. Therneau, Mayo Clinic
Elizabeth J. Atkinson, Mayo Clinic
Kellie J. Archer, Virginia Commonwealth University
Jeanette E. Eckel-Passow, Mayo Clinic

Batch effects are defined as probe-specific systematic non-biological variation between groups of samples (batches) due to experimental features of the experiment that are not corrected by global normalization methods (e.g. quantile, loess). Principal components analysis (PCA) is commonly used to determine if batch effects exist after applying a global normalization method. However PCA will identify factors that contribute maximum variance and thus will not necessarily detect batch effects. We present an extension of principal components analysis (PCA) to visualize and quantify the existence of batch effects, called guided PCA (gPCA). We apply gPCA to two copy number variation case studies: the first study consisted of 614 samples from a breast cancer family study using Illumina Human 660 bead-chip arrays whereas the second case study consisted of 703 samples from a family blood pressure study that used Affymetrix SNP Array 6.0. We show how gPCA can be used to determine the percent of variability that is specific to batch. Although we will show examples using copy number data, gPCA can be used on other data types as well.

email: reesese@vcu.edu

3p. DISTRIBUTION OF ALLELE FREQUENCIES AND EFFECT SIZES AND THEIR INTERRELATIONSHIPS FOR COMMON GENETIC SUSCEPTIBILITY VARIANTS

Ju-Hyun Park, National Cancer Institute, National Institutes of Health*
Mitchell H. Gail, National Cancer Institute, National Institutes of Health
Clarice R. Weinberg, National Institute of Environmental Health Sciences, National Institutes of Health
Raymond J. Carroll, Texas A&M University
Charles C. Chung, National Cancer Institute, National Institutes of Health
Zhaoming Wang, National Cancer Institute, National Institutes of Health
Stephen J. Chanock, National Cancer Institute, National Institutes of Health
Joseph F. Fraumeni, National Cancer Institute, National Institutes of Health
Nilanjan Chatterjee, National Cancer Institute, National Institutes of Health

Recent discoveries from genome-wide association studies provide a unique opportunity to examine population genetic models for complex traits. In this report, we investigate distributions of various population genetic parameters and their interrelationships using estimates of allele frequencies and effect-size parameters for ~ 400 susceptibility SNPs across qualitative and quantitative traits. We calibrate our analysis by statistical power for detection of SNPs to account for overrepresentation of variants with larger effect sizes in currently known SNPs that are expected due to statistical power for discovery. Across all traits, an inverse relationship existed between regression effects and allele frequencies. This trend was remarkably strong for type I diabetes, a trait that is most likely to be influenced by selection, but was modest for other traits such as human height or late-onset diseases. For most traits, the set of less common SNPs (5-20%) contained an unusually small number of susceptibility loci and explained a relatively small fraction of heritability compared with what would be expected from the distribution of SNPs in the general population. These trends could have several implications for future studies of common and uncommon variants.

email: parkj3@mail.nih.gov

3q. SELECTING A STATISTICAL TEST TO DETECT ASSOCIATIONS WITH GROUPS OF GENETIC VARIANTS: A USER'S GUIDE

John Ferguson, Yale University
Hongyu Zhao, Yale University
William Wheeler, IMS
Yi-Ping Fu, National Cancer Institute, National Institutes of Health
Ludmila Prokunina-Olsson, National Cancer Institute, National Institutes of Health
Joshua N. Sampson, National Cancer Institute, National Institutes of Health*

We show that the majority of statistics that test for an association between a group of genetic variants and a phenotype share a common framework. Within this simple and intuitive framework, we then show that investigators can easily identify those statistics that are best suited for their specific studies. The performance of these statistics depends on properties such as the proportion of SNPs associated with the disease, the direction of effect, and the relationship between effect size and minor allele frequency (MAF). Consequently, no single test is universally preferable. Furthermore, this framework enables investigators to tailor the standard test statistics to the particulars of their own problem, as we demonstrate by designing a shrinkage-based estimator for a study of bladder cancer. Although studies have found that the UGT1A cluster on chromosome 2q37 is associated with multiple cancers, we demonstrate that its entire relationship with bladder cancer appears to be mediated through a single SNP.

email: joshua.sampson@nih.gov

3r. A NEW PENALIZED REGRESSION APPROACH TO TESTING QUANTITATIVE TRAIT-RARE VARIANT ASSOCIATION

Sunkyung Kim, University of Minnesota*
Wei Pan, University of Minnesota
Xiaotong Shen, University of Minnesota

In statistical data analysis, penalized regression has been considered as an attractive approach for its ability for simultaneous dimension reduction and parameter estimation. Though penalized regression methods have shown many advantages in variable selection and outcome prediction over other approaches, there is a paucity of literature on its application to hypothesis testing, e.g in genetic association analysis. In this study, to test for association between a quantitative trait and a group of rare variants in a candidate gene, we apply a new penalized regression method with a group Truncated L1-penalty (TLP), which groups predictors in a data-adaptive way. Its performance is compared with some existing tests via simulations. Simulation studies suggest that overall the method is competitive, and its estimation error is compared favorably to that of ordinary least squares (OLS). Finally, we apply the method to real sequence data from the Genetic Analysis Workshop 17 (GAW17).

email: kimx1606@umn.edu

3s. INCORPORATING HETEROGENEITY INTO META-ANALYSIS OF GENOMIC DATA: A WEIGHTED HYPOTHESIS TESTING APPROACH

Yihan Li, Pennsylvania State University*
Debashis Ghosh, Pennsylvania State University

There is now a large literature on statistical methods for the meta-analysis of genomic data from multiple studies. However, a crucial assumption for performing analyses is that the data exhibit small between-study variation. Procedures for addressing this assumption are currently not available. In this article, we exploit a weighted hypothesis testing framework proposed by Genovese et al. (2006, Biometrika 93, 506 - 524) to incorporate tests of between-study variation in the meta-analysis context. Several weighting schemes are considered and compared using simulation studies. In addition, we illustrate application of the proposed methodology using data from several high-profile stem cell gene expression datasets.

email: yihanli@psu.edu

3t. MODELING HAPLOTYPE EFFECTS IN A GENETIC REFERENCE POPULATION: A BAYESIAN COLLABORATIVE CROSS TOOLKIT

Zhaojun Zhang, University of North Carolina at Chapel Hill*
Wei Wang, University of North Carolina at Chapel Hill
William Valdar, University of North Carolina at Chapel Hill

In model organism genetics there has been an increasingly focus on developing Genetic Reference Populations (GRPs). GRPs are genetically diverse populations suitable for gene mapping that are 1) designed to facilitate replicable studies, and 2) are genetically derived from a smaller set of well-characterized founder populations. Because of their common ancestry, each GRP individual's genome can be decomposed into a mosaic of founder haplotypes (ie, genome segments). We develop a Bayesian hierarchical model for estimating QTL effects in GRPs, focusing on an emerging mouse GRP, the Collaborative Cross (CC). Our modeling framework seeks to characterize the effect of genetic variation at a given genomic location on continuous or dichotomous phenotypic outcomes. Importantly, rather than modeling affects of particular SNPs, we use a more inclusive and general model that is based on effects that differ across founder haplotypes, these haplotypes being known with uncertainty and inferred using a Hidden Markov model applied to genotype or sequence data. We demonstrate the use of our method for characterizing posterior densities of QTL effects on real continuous and dichotomous data. We also show how further exploration of haplotypic effects is aided by a dirichlet process model.

email: zzj@cs.unc.edu

3u. MULTI-MARKER ASSOCIATION ANALYSIS WITH MULTIPLE PHENOTYPES IN FAMILIES

Yiwei Zhang, University of Minnesota*
Saonli Basu, University of Minnesota

Studying the joint effect of multiple genetic variants on multiple correlated traits has gained great impetus in recent days. Genetic variants can have influence on all these phenotypes and modeling these phenotypes jointly can significantly improve power for detection of these variants. Furthermore, pedigree data provides more information on multivariate phenotype association analysis; however the traditional methods for multivariate modeling impose significant computational challenges. Here under a rapid feasible generalized least square framework, we compare five approaches such as minimum p-value method (minP), canonical correlation method (CCA), linear mixed model (LME), Fisher's method and O'brien's test through extensive simulation studies. We propose computationally efficient ways of applying these methods to family data in order to study the effects of multiple variants on multiple traits and show under what scenarios we gain power by modeling these traits simultaneously. We show that if the traits are partially associated with a variant, CCA works much better than other methods. However, if all the traits are associated with the variant, minP works surprisingly well. We have also studied their performance on Minnesota Center for Twin and Family Research data, a longitudinal genome-wide study on correlated behavioral traits.

email: zhang_ivy_1@hotmail.com

3v. MULTI-STAGE SEQUENCE/IMPUTATION DESIGN

Thomas J. Hoffmann, University of California San Francisco*
John S. Witte, University of California San Francisco

Recent findings suggest that rare variants play an important role in disease, and studies have moved to studying rare variants to explain more of the missing genetic heritability. There are several different study designs for rare variants. One option would be to sequence everyone. On the other extreme, one could genotype everyone on arrays, and then impute everyone using existing sequence data (e.g., the 1000 Genomes Project). However, if disease-causing rare variants are unique to a certain population, then one needs to directly sequence individuals from that population to detect these rare variants. Although sequencing costs are decreasing, it is also generally not cost-effective to sequence everyone, unless one is studying a rare, highly penetrant diseases, e.g., in a family, as there will be little power. An alternative design is to compromise with a multi-stage approach in which one sequences a subset of individuals, and use these data to impute into the rest of the subjects genotyped on an array. There are many challenging design issues we consider here, such as how many subjects should be sequenced to capture certain ranges of rare allele frequency variants, which individuals should be sequenced, and the challenges with imputation.

email: tjhoffm@gmail.com

3w. AN APPLICATION OF THE PROPORTIONAL ODDS MODEL TO GENETIC ASSOCIATION STUDIES

Kai Wang, University of Iowa*

Powerful statistical methods are essential to successful genetic association studies. To achieve high power, an inheritance model is often assumed for the trait being studied. To ameliorate issues accompanied by model mis-specification, one alternative is to use the optimal p-value resulted from several inheritance models. However, it is none-trivial to assess the significance of this optimal p-value. In the current research, we switch the role of genotype and phenotype in usual regression analysis. That is, the genotype is used as the response variable and the phenotype the explanatory variable. The proportional odds model is then used by simply assuming the existence of an ordering in the three genotypes at a single nucleotide polymorphism. This method

encompasses a wide range of inheritance models and yet avoids complicated asymptotic distribution for the test statistic. Such a conditioning on phenotype approach provides a unified treatment of continuous phenotype, dichotomous phenotype, or selected phenotype. Extensive simulation studies are conducted to assess its performance. It is also applied to Genetic Analysis Workshop 16 rheumatoid arthritis data.

email: kai-wang@uiowa.edu

3x. AN EMPIRICAL EVALUATION OF ARRAY NORMALIZATION FOR AGILENT microRNA EXPRESSION ARRAYS

Li-Xuan Qin, Memorial Sloan-Kettering Cancer Center*
Qin Zhou, Memorial Sloan-Kettering Cancer Center
Jaya Satagopan, Memorial Sloan-Kettering Cancer Center
Samuel Singer, Memorial Sloan-Kettering Cancer Center

Methods for array normalization have been developed for mRNA expression arrays, such as median normalization and quantile normalization. These methods assume few or symmetric differential expression of markers on the array. The performance of the existing normalization methods need to be re-evaluated when applied to microRNA arrays, which consist of a few hundred markers and a reasonable fraction of them are anticipated to have disease-relevance. We empirically examined sources of variations in miRNA array data using a set of Agilent arrays in liposarcoma (n=56) and evaluated normalization methods using Solexa sequence data on a subset of these tumors (n=29) as the gold standard. We found that there is minimum variation between replicate probes for the same target sequence and moderate variation between multiple target sequences for the same miRNA. There is moderately high correlation between Agilent data and Solexa data. Quantile normalization has slightly improved the correlation with Solexa data, as well as the detection of differentially expressed microRNAs both in terms of statistical significance and the direction of change.

email: qinl@mskcc.org

3y. ASSOCIATION ANALYSIS OF RARE VARIANTS WITH INCOMPLETE GENETICS DATA

Yijuan Hu, Emory University*
Danyu Lin, University of North Carolina at Chapel Hill

Biological evidence suggests that rare variants account for a large proportion of the genetic contributions to complex human diseases. Although technological advances in high-throughput sequencing platforms have made it possible to study rare variants, it is economically infeasible to sequence a large number of subjects. Many studies selected only a subset of subjects for sequencing from a large cohort, which typically has GWAS data. The current strategy is to impute individual variants based on the joint haplotype distributions of observed and missing variants. Unfortunately, rare variants cannot be imputed reliably. Thus, we propose to impute the burden scores rather than individual variants. In addition, we integrate the imputation of rare variants and the association analysis into a single likelihood framework so as to make proper inference. We rigorously establish the asymptotic properties of the proposed estimators and develop efficient and stable numerical algorithms that are scalable to genome-wide analysis. The common practice of single imputation, which treats the imputed genetic values as known in downstream association analysis, generally produces inflated type I error. We conduct extensive simulation studies to compare the maximum likelihood and imputation approaches under various scenarios. We provide an illustration with the Exome Sequencing Project (ESP) data.

email: yijuan.hu@emory.edu

3z. A WEIGHTED AVERAGE LIKELIHOOD RATIO TEST WITH APPLICATION TO RNA-Seq DATA

Yaqing Si, Iowa State University*
Peng Liu, Iowa State University

The recent RNA-seq technology has been widely adopted as an attractive alternative to microarray-based methods. One of the most important questions in RNA-seq experiments is to compare gene expressions from different treatment groups. Although several analyzing tools have been published, there is no theoretical justification whether these methods are optimal or how to search for the optimal test. We proposed a weighted-average likelihood ratio (WALR) test to address this question. The WALR test is constructed through an Empirical Bayes approach based on probability models, such as Poisson and Negative Binomial distributions. We show that our test has higher testing power than other compared methods, including the popularly applied edgeR, DESeq and baySeq methods. We also utilize the statistics of WALR tests to control false discovery rate (FDR) for both WALR test and any other test. Simulation results show that the FDR estimation is close to the true values and better than current methods in practice. Furthermore, instead of focusing on testing gene that are differentially expressed (DE), our test can be used to investigate more general questions, for instance, detecting genes that have fold changes (FC) exceeding a threshold, or testing genes with the expressions in one sample greater than in the other.

email: siyaqing@iastate.edu

3aa. ORDER STATISTIC FOR ROBUST GENOMIC META-ANALYSIS

Chi Song, University of Pittsburgh*
George C. Tseng, University of Pittsburgh

Meta-analysis techniques have been widely extended and applied in genomic applications, especially for combining multiple transcriptomic studies. In this paper, we proposed an order statistic of p-values (rth order p-value, rOP) across combined studies as the test statistic. We illustrated different hypothesis settings that detect gene markers differentially expressed (DE) in all studies, in majority of studies, or in one or more studies, and specified rOP as a suitable method for detecting DE genes in majority of studies. Theoretically, rOP was found connected to naive vote counting method and can be viewed as a generalized form of vote counting with better statistical properties. We developed methods to estimate the parameter r in rOP for real applications. Statistical properties such as its behavior and one-sided testing correction were explored. Power calculation showed better performance of rOP compared to classical Fisher's method, Stouffer's method, minimum p-value method and maximum p-value method under the focused hypothesis setting. The method was applied to three real meta-analysis examples including major depressive disorder, brain cancer and diabetes. The results demonstrated that rOP provides a more generalizable, robust and sensitive statistical framework to detect disease related markers.

email: chs108@pitt.edu

3ab. ASYMPTOTIC PROPERTIES AND CONVERGENCE RATE IN SOLVING MODELS OF LINKAGE DISEQUILIBRIUM MAPPING

Jiangtao Luo, University of Nebraska Medical Center*

In genetic mapping of complex traits using molecular markers, we often face a challenge of solving a mixture model with multiple components specified by a frequency structure different from ordinary mixture models. We describe an EM version algorithm to solve this special form of mixture, showing great power for gene detection. We prove that the algorithm obey a Rao-Blackwellization process. These estimators are not only asymptotically consistent, but also have less standard errors. We further prove a capture theorem and quadratic convergence of the algorithm. Computer simulation is used to test the algorithm, with results supporting our theoretical findings. The algorithm is used to analyze genetic data from an obesity genetic project, validating its practical usefulness and utilization.

email: jiangtao.luo@unmc.edu

4. CLINICAL TRIALS/BIOPHARMACEUTICALS/ MEDICAL DEVICES

4a. A PHASE II FACTORIAL DESIGN FOR COMBINATION CODEVELOPMENT IN ONCOLOGY BASED ON PROBABILITY OF CORRECT SELECTION

Xinyu Tang, University of Arkansas for Medical Sciences*
William Mietlowski, Novartis Oncology

The United States Food and Drug Administration (FDA) (2010) has recently issued a draft guidance entitled "Guidance for industry -codevelopment of two or more unmarketed investigational drugs for use in combination." In the case where each of the two investigational drugs has some activity and both can be given individually, the guidance recommends that a four-arm Phase II factorial design should be used to compare the combination of two investigational drugs to each drug alone and to placebo or standard of care. The aim of the design is to establish contribution of each component in Phase II without exposing a large number of patients to relatively ineffective treatments, enabling a simple two-arm randomized Phase III trial to be conducted. We used Green's criteria (Green 2006) for selecting the best treatment in a factorial design using a time-to-event endpoint as the basis of a Phase II strategy relying on the probability of correct selection concept applied to a progression-free survival endpoint. We claim that selection of the combination as the best agent using Green's criteria may provide evidence of component contribution in Phase II trials. We investigate the operating characteristics of this proposal using simulations.

email: xtang@uams.edu

4b. BAYESIAN APPLICATION FOR A CLINICAL TRIAL WITH CORRELATED CONTINUOUS AND BINARY OUTCOMES

Ross Bray, Baylor University*
John W Seaman Jr., Baylor University
James Stamey, Baylor University

We look at a Bayesian application of a correlated bivariate pharmaceutical model dealing with a binary safety variable and a continuous efficacy variable where the binary safety variable is given a Bernoulli distribution and the continuous efficacy variable is given a Normal distribution. We discuss the model specifications and choice of priors, and then implement the model on an example where the safety variable is suicide attempt, and the efficacy variable is weight loss. We finish by looking at a simulation to test model effectiveness.

email: ross_bray@baylor.edu

4c. SCALED BIOSIMILARITY MARGINS FOR HIGHER VARIABLE BIOLOGIC PRODUCTS

Nan Zhang, Amgen, Inc.
Jun Yang, Amgen, Inc.*
Shein-Chung Chow, Duke University
Eric Chi, Amgen, Inc.
Laszlo Endrenyi, University of Toronto

In FDA guidance (2001), the criteria for bioequivalence (BE) in small molecular drugs rely on fixed BE limits. However, in later literature the scaled BE limits (Endrenyi, 2011) were suggested for higher variable drugs. It is well known that biological drug products often have larger variation compared with small molecular drugs. Therefore, when assessing the biosimilarity of follow-on biologics, the choice of biosimilarity margins should take into consideration the variability. In this study, we explored the impact of variability on biosimilarity margins for the average biosimilarity assessment criterion. Based on the relationship between variability and biosimilarity margins, several scaled biosimilarity margins were proposed for highly variable biological drug products and their power and type I error were compared in various parameter settings.

email: juny@amgen.com

4d. DOSE FINDING DESIGNS FOR MODELING IMMUNOTHERAPY OUTCOMES: A PRACTICAL APPROACH IN A METASTATIC MELANOMA PHASE I TRIAL

Cody C. Chiuzan, Medical University of South Carolina*
Elizabeth Garrett-Mayer, Medical University of South Carolina

Immunotherapy offers a great promise as a new dimension in cancer treatment. This research is motivated by an ongoing project aimed to establish the recommended phase II dose and evaluate the biologic and immunologic parameters associated with the transfer of genetically engineered lymphocytes (T *cell receptor* (TCR) transduced T cells with low dose of IL-2) in melanoma patients. An advantage of this clinical approach relies on the unique sequences within the TCR that enable us to monitor the persistence as a continuous outcome. We propose an adaptive randomization design that models percent persistence as the main outcome. In the first stage, cohorts of two patients are to be treated at each of the five escalating doses. In the second stage, patients are adaptively randomized to doses based on the observed persistence from previous patients. Persistence is generated from a beta-binomial distribution, where patient heterogeneity is controlled by beta distribution. The presented simulations are based on models incorporating five different dose-response scenarios and three efficacy criteria. Based on our simulations, the adaptive model results in a larger number of patients assigned to doses with higher persistence, smaller bias, and increased precision in dose estimation, when compared to the equal allocation design.

email: chiuzan@musc.edu

4e. ELLIPTICAL LIKELIHOOD RATIO TEST FOR HOMOGENEITY OF ORDERED MEANS

Xiao Zhang, University of Rochester*
Michael P. McDermott, University of Rochester

Order restrictions on model parameters arise in many practical problems, including that of testing equality of means from independent normal populations. The restricted parameter space in this setting can be expressed as a closed, convex pointed polyhedral cone. The most well-known and extensively studied approach to this problem is the likelihood ratio test, but computation of the test statistic and its null distribution limit its use. Based on work by Pincus (1975), Akkerboom (1990) and Conaway et al. (1990) proposed the circular likelihood ratio test that approximates the restricted parameter space with a circular cone. The resulting test statistic and its null distribution are available in closed form, making the test easy to implement, but its operating characteristics depend on the quality of the approximation of the original pointed polyhedral cone with a circular cone. We develop an elliptical likelihood ratio test that approximates the restricted parameter space with a more flexible elliptical cone, which includes the circular cone as a special case. The operating characteristics of this test and issues regarding its implementation will be discussed.

email: xiao_zhang@urmc.rochester.edu

4f. ANALYSIS OF MULTIPLE NON-COMMENSURATE OUTCOMES IN PSYCHIATRY

Frank B. Yoon, Mathematica Policy Research, Inc.*
Garrett M. Fitzmaurice, Harvard School of Public Health
Stuart R. Lipsitz, Harvard Medical School
Nicholas J. Horton, Smith College
Sharon-Lise T. Normand, Harvard Medical School

Multiple outcomes in randomized and observational studies in psychiatry are often non-commensurate, for example, measured on different scales or constructs. Standard multiplicity adjustments can control for Type I error, though such procedures can be overly conservative when the outcomes are highly correlated. Recent literature demonstrates that joint tests can capitalize on

the correlation among the outcomes and are more powerful than univariate procedures using Bonferroni adjustments. However, joint tests are little used in practice, perhaps, due in part, to the specification of a joint model for the non-commensurate outcomes. Additionally, software routines to estimate joint models have not been widely publicized despite their wide availability. This work presents an evaluation of likelihood and quasi-likelihood methods for jointly testing treatment effects in a simulation study. Applications to a clinical trial and an observational study of mental health care illustrate their benefits. Adoption of these methods will lead to more efficient psychiatric clinical trials.

email: fyoon@mathematica-mpr.com

4g. EVALUATION FOR TIME TO ONSET OF DRUG ACTION

Ziwen Wei, University of Connecticut*
Luyan Dai, Boehringer Ingelheim
Naitee Ting, Boehringer Ingelheim

Time to onset of drug action is an important, but not well-studied problem. In most of clinical trials where time to onset is assessed, repeated statistical hypothesis tests are performed. However, in addition to hypothesis testing, this problem can also be viewed from an estimation point of view. On this basis, this manuscript proposes to apply a modeling approach to help estimate time to onset of drug action. Simulations are performed to evaluate the properties of the proposed approach, and an example of anti-hypertensive study is presented to illustrate how the proposed method can be applied to a real world case.

email: ziwenwei@hotmail.com

4h. META-ANALYSIS OF ONE OUTCOME FROM GROUP SEQUENTIAL TRIALS WITH COMPOSITE OUTCOMES: ARE STANDARD METHODS APPROPRIATE?

Abigail B. Shoben, The Ohio State University*

Composite outcomes are the primary outcome in many clinical trials. A composite outcome combines multiple events into one outcome, such as the outcome of death or tumor progression in a cancer clinical trial. These composite outcomes are convenient because they provide a way to collapse over competing risks and provide simple interpretation to physicians and others evaluating effectiveness. However, composite outcomes may not provide clarity if treatment effects differ between the outcomes being combined and attempts to separate the effects post-hoc either in a single trial or in combined meta-analyses may be problematic. An additional complication is that data from randomized trials are monitored for safety and often employ sequential analysis methods. Such safety and sequential monitoring may be done using the composite outcome or using one or more outcomes separately. Differential monitoring of one or more outcomes in the composite outcome may further complicate post-hoc meta-analyses. We consider as an example recent meta-analyses of data from randomized trials of magnesium sulfate on neurological deficits and death in preterm infants. We illustrate the problems resulting from the separation of the composite outcome into the separate outcomes in these meta-analyses and provide guidance for future studies.

email: ashoben@cph.osu.edu

4i. STATISTICAL ANALYSIS OF EVALUATING THE CLINICAL UTILITY OF QUANTITATIVE REAL-TIME LOOP-MEDIATED ISOTHERMAL AMPLIFICATION FOR DIAGNOSIS OF LOWER RESPIRATORY TRACT INFECTIONS

Peng Zhang, Peking University*
Peichao Peng, Peking University
Yu Kang, Peking University
Minping Qian, Peking University

It is challenging to evaluate the clinical utility of qrt-LAMP for diagnosis of lower respiratory tract infections (LRTI), which lacks good diagnostic methods. Partially, difficulties lie in that multiple pathogens co-exist in the host system, where some pathogens colonize without causing problems and some pathogens rapidly grow and cause infection. Clinically, it is important to differentiate colonization from infection, and then recommend the right choice of antibiotic treatments. New statistical methods are needed to tackle such problems. We first utilize zero-inflated mixture models to estimate prevalence of pathogens through LAMP, and demonstrate that LAMP shows consistency with the results from the standard culture methods. We then use zero-inflated Tobit model to adjust for baseline covariates on both probabilities of carrying pathogens and quantities of pathogens for carriers. With clear clinical interpretations of such results, it furthers validates accuracy of LAMP measurements from experiments. Finally, we novelly design U-scores which incorporate both absolute quantities of pathogens and their symbiosis information. Changing-point detection methods clearly reveal two change points for these U-scores, which correspond to three phases in the biological growth of bacteria. We use piecewise-linear regression to identify such changing-points.

email: pczhang@pku.edu.cn

4j. MONOTONICITY ASSUMPTIONS FOR EXACT UNCONDITIONAL TESTS IN BINARY MATCHED-PAIRS DESIGNS

Xiaochun Li, New York University*
Mengling Liu, New York University
Judith D. Goldberg, New York University

We present the calculation of exact p-values (M, C and E +M) and their corresponding exact sizes for a noninferiority test for a 2 x 2 matched pairs design. Our research is motivated by a trial in cervical cancer that was designed to test the hypothesis that the sensitivity and specificity of the experimental treatment (Dart) were not worse than (non-inferior to) the sensitivity and specificity of the reference control (Colposcopy). The existing computation of the three p-values (M, C and E +M) and their exact sizes generally requires that the Barnard Convexity Condition (BCC) be satisfied; this can be challenging to prove theoretically and sometimes has to rely on numerical verification. By a simple reformulation, we show that a weaker condition, Conditional Monotonicity Condition (CMC), is sufficient to calculate all three p-values and their exact sizes without the need to use numerical verification. Moreover, the CMC is applicable to both noninferiority tests and superiority tests.

email: xiaochunlee@gmail.com

4k. A PHASE I/II CLINICAL TRIAL FOR DRUG COMBINATIONS

Beibei Guo, University of Texas MD Anderson Cancer Center*
Yisheng Li, University of Texas MD Anderson Cancer Center

Traditional clinical trials evaluate a single agent at a time. A phase I trial is designed to establish the maximum tolerated dose and a phase II trial is used to determine whether the agent is sufficiently promising relative to a standard treatment. We consider combinations of two agents and combine the two phases into one. We present two Bayesian approaches that jointly model the probabilities of toxicity and efficacy based on some pre-specified order constraints between dose combinations. The first one is based on Bayesian isotonic transformation, and the second one is based on stochastic ordering, which makes less extreme ordering. Based on the joint posterior probabilities, we develop a dose combination finding algorithm that assigns patients to the current best dose combination under certain criterion. At the end the trial, the optimal combination dose is selected. The performances of the two proposed methods are examined through simulation studies. The results show that the method based on stochastic ordering leads to better allocations, selects the target dose combination more frequently, and is less aggressive.

email: beibeiguo@gmail.com

4l. OVERVIEW OF STATISTICAL ISSUES IN THE ANALYSIS OF CONTINUOUS GLUCOSE MONITORING

Chava Zibman, U.S. Food and Drug Administration*

The FDA is currently in the process of developing guidance documents on the clinical evaluation of continuous glucose monitors (CGMs) and artificial pancreas systems. This presentation will examine some of the statistical issues related to these products. I will begin with a discussion of the metrics used in the analysis of CGM performance. In particular, I will discuss methods used to adjust for autocorrelation when calculating the sample size for CGM studies, given study design and choice of endpoint. Continuing with a focus on the time-series nature of glucose monitoring data, I will also address the evaluation of accuracy over time and rate-of-change as addressed in recent works by Kovatchev and others.

email: chava.zibman@fda.hhs.gov

4m. BIOLOGICAL OPTIMUM DOSE FINDING FOR NOVEL TARGETED AGENTS

Hao Liu, Baylor College of Medicine*

Biological targeted agents can generate optimal treatment effect at a level well below the maximum tolerated dose. Toxicity-guided dose selection designs thus may not work for identifying the optimal dose level. In this article, we propose a Bayesian adaptive design for dose-finding trials guided by biological responses that directly measure the targeted effect of the study agent. A subject will receive the treatment at the dose level that has the best chance for improving biological response, provided that the dose level has limited toxicity. In addition, a Bayesian rule is incorporated to stop a trial early based on the predictive probability of adverse events. Simulation studies show that the method performed reasonably well. We illustrate the method using an example of a dose-finding trial in advanced renal cell carcinoma.

email: haol@bcm.edu

4n. HIERARCHICAL BAYESIAN METHODS FOR COMBINING EFFICACY AND SAFETY IN MULTIPLE TREATMENT COMPARISONS

Hwanhee Hong, University of Minnesota*
Bradley P. Carlin, University of Minnesota

Biomedical decision makers confronted with questions about the comparative effectiveness and safety of interventions, often wish to combine all sources of data. Such multiple treatment comparisons (MTCs) may or may not include head-to-head randomized controlled trials of the treatments of primary interest, instead relying largely on indirect comparisons (say, trials that separately compare each treatment to placebo). In such settings, hierarchical Bayes-MCMC meta-analytic methods offer

a natural approach (e.g., by enabling full posterior inference on the probability that a given treatment is best). In this paper, we summarize the current state of such methods in the binary response setting, and consider extension to the case of multiple outcomes (say, on both efficacy and safety) where we account for correlation and use clinically-informed weights to arrive at an overall decision regarding the best treatment. We illustrate our methods with data from a recent MTC, comparing pharmacological treatments for female urinary incontinence. We also offer several simulations to support the use of our methods over more standard approaches that ignore cross-endpoint correlation. We close with a discussion of our results and a few avenues for future methodological development.

email: hong0362@umn.edu

5. COMPUTATIONALLY INTENSIVE METHODS / HIGH DIMENSIONAL DATA

5a. GLOBAL HYPOTHESIS TESTING FOR HIGH DIMENSIONAL REPEATED MEASURES OUTCOMES

Yueh-Yun Chi, University of Florida*
Matthew Gribbin, Human Genome Sciences
Lamers Yvonne, University of British Columbia
Jesse F. Gregory, University of Florida
Keith E. Muller, University of Florida

High-throughput technology in metabolomics, genomics, and proteomics gives rise to high dimension, low sample size data when the number of metabolites, genes, or proteins exceeds the sample size. For a limited class of designs, the classic univariate approach for Gaussian repeated measures can provide a reasonable global hypothesis test. We derive new tests that not only accurately allow more variables than subjects, but also give valid analyses for data with complex between- and within-subject designs. Our derivations capitalize on the dual of the error covariance matrix, which is non-singular when the number of variables exceeds the sample size, to ensure correct statistical inference and enhance computational efficiency. Simulation studies demonstrate that the new tests accurately control Type I error rate and have reasonable power even with a handful of subjects and a thousand outcome variables. We apply the new methods to the study of metabolic consequences of vitamin B6 deficiency. Free software implementing the new methods applies to a wide range of designs, including one group pre- and post-intervention comparisons, multiple parallel group comparisons with one-way or factorial designs, and the adjustment and evaluation of covariate effects.

email: yychi@biostat.ufl.edu

5b. AN EXAMPLE OF USING SWEAVE TO CREATE AND MAINTAIN A LARGE, DYNAMIC STATISTICAL REPORT: PREVALENCE AND EFFECTS OF POTENTIALLY DISTRACTING NON-CARE ACTIVITIES DURING ANESTHESIA CARE

David Afshartous, Vanderbilt University
Steve Ampah, Vanderbilt University*
Jason Slage, Vanderbilt University
Eric Porterfield, Vanderbilt University
Samuel K. Nwosu, Vanderbilt University

Sweave is a useful tool that facilitates the production of reproducible research via the embedding of R code in LaTeX documents that generate statistical reports. The main advantage is that the reports are dynamic in the sense that they can be easily updated if data or analysis change. However, the ease with which the report is updated is dependent upon both the type of data change and the chosen structure of the R code and Sweave document. We provide an example of an analysis of a large and complex data set using Sweave where there were changes to various aspects of the data. Specifically, we provide recommendations for managing and increasing the efficiency of report updates in response to: 1) new observations, 2) new variables, and 3) new categories for existing variables. The data are from a study of the prevalence of potentially distracting non-care activities and their effects on vigilance and workload during anesthesia care.

email: d.afshartous@vanderbilt.edu

5c. ITERATIVELY REWEIGHTED POISSON REGRESSION FOR FITTING GENERALIZED LINEAR MODEL WITH MULTIPLE RESPONSES

Yiwen Zhang, North Carolina State University*
Hua Zhou, North Carolina State University

Generalized linear models with multiple responses (MGLMs) are seeing wider use in modern applications such as pattern recognition, document clustering, and image reconstruction. Examples of MGLMs include multinomial-logit models, Dirichlet-multinomial overdispersion models, and negative-multinomial models. Maximum likelihood estimation of MGLMs is difficult due to the high-dimensionality of the parameter space and possible non-concavity of the log-likelihood function. In this article, we propose iteratively reweighted Poisson regression as a unified framework for maximum likelihood estimation of MGLMs. The derivation hinges on the minorization-maximization (MM) principle which generalizes the celebrated expectation-maximization (EM) algorithm. MM algorithm operates by constructing a surrogate function with parameters separated. Optimizing such a surrogate function drives the objective function in the correct direction. This leads to a stable algorithm which possesses good global convergence property and is extremely simple to code. The proposed algorithm is tested on classical and modern examples.

email: yzhang31@ncsu.edu

5d. JOINT ESTIMATION OF MULTIPLE PRECISION MATRICES

T. Tony Cai, University of Pennsylvania
Hongzhe Li, University of Pennsylvania
Weidong Liu, Shanghai Jiao Tong University
Jichun Xie, Temple University*

Precision matrices reveal conditional dependency structures between multivariate normal random variables. Some heterogenous data share similar dependency structures among different groups. One of the examples is the multiple-tissue gene expression data. Evidence has shown that the gene regulatory networks share similarities across different tissues and yet have tissue-specific interaction mechanism. To improve the estimation efficiency, we develop a joint estimation method to infer precision matrices. Theoretical and Numerical studies show that our method has improved support recovery results compared with separate estimation methods and other competing joint estimation methods. We further apply our method to the mouse gene expression data and obtain the gene regulatory networks for multiple mouse tissues.

email: jichun@temple.edu

5e. HOW TO BOOTSTRAP fMRI DATA?

Sanne Roels, Ghent University, Belgium*
Tom Loeys, Ghent University, Belgium
Beatrijs Moerkerke, Ghent University, Belgium

Over the last decade the bootstrap procedure is gaining popularity in the statistical analysis of neuro-imaging data. As fMRI data are complexly structured with both temporal and spatial dependencies, such bootstrap procedures may indeed offer an elegant and powerful solution. We provide a thorough investigation on the most appropriate bootstrapping procedure for fMRI data. Friman and Westin (2005, NeuroImage) showed that a bootstrap procedure based on pre-whitening the temporal structure of fMRI time series is superior to procedures based on wavelets or Fourier decomposition of the signal, especially in the case of blocked fMRI designs. Several bootstrap schemes can be exploited though for the re-sampling of residuals from a fitted general linear model. We examine here the differences between 1) bootstrapping pre-whitened residuals which are based on parametric assumptions of the temporal structure, 2) a blocked bootstrapping which avoids making such assumptions (with several variants like the circular bootstrap,...), and 3) a combination of both bootstrap procedures. Because the dependence in the residuals of a misspecified GLM might differ in blocked and event-related fMRI designs we investigated the bootstrapping schemes for both design types. Based on real data and simulation studies, we discuss the temporal and spatial properties of the different bootstrap procedures.

email: sanne.roels@ugent.be

5f. THE HOSMER-LEMESHOW GOODNESS OF FIT TEST: DOES THE GROUPING REALLY MATTER?

Hillary M. Rivera, University of Arkansas for Medical Sciences*
Zoran Bursac, University of Arkansas for Medical Sciences
D. Keith Williams, University of Arkansas for Medical Sciences

The Hosmer and Lemeshow goodness-of-fit statistic (HLGOF) is a well established method of assessing model fit for logistic regression. This statistic is based on the grouping of predicted probabilities. It is available in many software packages and routines that offer logistic regression modeling. Some suggest that the performance of HLGOF depends on the decile grouping that the test statistic is based on. We propose a modification to HLGOF decile risk approach via random grouping method. In this project we describe our approach and discuss its performance through simulations.

email: hillary.rivera@yahoo.com

5g. A BAYESIAN NON-PARAMETRIC POTTS MODEL WITH fMRI APPLICATION

Timothy D. Johnson, University of Michigan
Zhuqing Liu, University of Michigan*
Thomas E. Nichols, University of Warwick

We present a Bayesian non-parametric Potts model and use it to find activated, null and deactivated regions in single subject fMRI data for pre-surgical planning. While fMRI data is usually smoothed and statistic images threshold to strictly control false positive risk, in pre-surgical planning the highest possible resolution data is needed and false negatives pose a greater concern. Conditional on class membership, the intra-class voxels are independently distributed as a mixture of Dirichlet process priors mode. We modify the standard Gibbs potential function to include the marginal (over space) probability of class membership, on which we place a hyperprior distribution, allowing more flexibility with respect to prior beliefs about class membership and further allow the data to drive the posterior. The spatial regularization parameter is random as well and we estimate it in the posterior along with all other parameters. We use the Swendsen-Wang algorithm to estimate the posterior class membership probabilities and a decision theoretic framework to determine final class membership. We demonstrate our model on an fMRI data set used for pre-surgical planning and compare the results with a parametric Potts model. Our non-parametric model tends to give much sharper decision boundaries than the parametric version.

email: zhuqingl@umich.edu

5h. ORACLE INEQUALITIES FOR THE HIGH-DIMENSIONAL COX REGRESSION MODEL VIA LASSO

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

We consider the non-asymptotic properties of the regularized high-dimensional Cox regression via lasso. Existing literature focuses on linear models and generalized linear models with Lipschitz loss functions, where the empirical risk functions are the summations of independent and identically distributed (iid) losses. The (negative) partial likelihood function for censored survival data, however, is neither a sum of iid terms nor Lipschitz. We first approximate the partial likelihood function by a sum of iid terms that are not Lipschitz, and then provide the non-asymptotic oracle inequalities for the penalized partial likelihood function with lasso penalty.

email: kongsc@umich.edu

5i. INTEGRATED MACHINE LEARNING APPROACH AS A TOOL FOR TESTING SNP-SNP INTERACTIONS

Hui-Yi Lin, H. Lee Moffitt Cancer Center & Research Institute*

A growing number of studies evaluate SNP-SNP interactions to compliment the irreproducible findings of individual single nucleotide polymorphisms (SNPs) associated with complex diseases. In this study, we evaluated an integrated method that combines two machine learning methods - Random Forests (RF) and Multivariate Adaptive Regression Splines (MARS). In this two-stage RF-MARS (TRM) approach, RF is first applied to detect a predictive subset of SNPs and then MARS is used as the second step to identify interaction patterns among the selected SNPs. RF variable selection was based on two approaches: out-of-bag classification error rate (OOB) and variable important spectrum (IS). We evaluated this integrated machine learning approach using three simulated models which contained one 2-way interaction in a data set with 1,000 subjects. Our results support that RFOOB had better performance than MARS and RFIS in detecting important variables. This study demonstrates that TRMOOB is more powerful in identifying SNP-SNP interaction patterns in a scenario of 100 candidate SNPs than the other two methods. This TRMOOB approach was also applied for identifying interactions among 2,653 SNPs associated with prostate cancer aggressiveness. The results suggest using the TRM approach can successfully identify interaction patterns in studies with a large number of SNPs.

email: hui-yi.lin@moffitt.org

5j. OPTIMAL MULTI-STAGE SINGLE-ARM PHASE II DESIGN BASED ON SIMULATED ANNEALING

Nan Chen, University of Texas MD Anderson Cancer Center*
J. Jack Lee, University of Texas MD Anderson Cancer Center

Simon's two stage design and other multi-stage designs are commonly used in phase II single-arm clinical trials because of its simplicity and small expected sample size under the null hypothesis. Solutions for most of these designs are based on the exhaustive search method. However, exhaustive search is very difficult to be extended to the high dimensional, multi-stage interim analysis. In this study, we propose a new simulated annealing (SA) method to optimize the early stopping boundaries which minimize the expected sample size. The SA method successfully reproduces the early stopping boundaries for the commonly used designs. For the multi-stage designs, we compare the results from the SA method the posterior credible interval method, the predictive probability method, and the decision-theoretic method. The SA method has the smallest expected sample sizes in almost all scenarios. The expected sample sizes from the SA method are generally 10-20% smaller than the results from the posterior credible interval method or the predictive probability method and are close to the decision-theoretical method with the loss function to control both type I and type II errors.

email: nchen2@mdanderson.org

5k. SuBLIME: SUBTRACTION-BASED LOGISTIC INFERENCE FOR MODELING AND ESTIMATION

Elizabeth M. Sweeney, Johns Hopkins University and National Institute of Neurological Disorders and Stroke, National Institutes of Health*
Russell T. Shinohara, Johns Hopkins University and National Institute of Neurological Disorders and Stroke, National Institutes of Health
Colin D. Shea, National Institute of Neurological Disorders and Stroke, National Institutes of Health
Daniel S. Reich, Johns Hopkins University and National Institute of Neurological Disorders and Stroke, National Institutes of Health
Ciprian M. Crainiceanu, Johns Hopkins University

Detecting incident and enlarging lesions is essential in monitoring progression of multiple sclerosis (MS). In clinical trials, lesion activity is observed by manually segmenting serial T2-weighted magnetic resonance images (MRI), which is time consuming and costly. Subtracting images from consecutive time points cancels stable lesions, leaving only new lesion activity. Manually segmenting subtraction images depicts a higher number of active lesions with greater inter-observer agreement. We propose Subtraction-Based Logistic Inference for Modeling and Estimation (SuBLIME), an automated method for segmenting voxel-level lesion incidence. Logistic regression models using various modalities of MRI from consecutive studies assign probabilities of lesion incidence. In this analysis, 110 MRI studies were analyzed; 11 studies from 10 patients with T1-weighted, T2-weighted, FLAIR

and PD volume acquisitions. With SuBLIME, on a voxel-level, lesion incidence can be detected with an area under the receiver operator characteristic curve of 99%. SuBLIME can also be applied using only the T2-weighted subtraction image, detecting lesion incidence with an area under the receiver operator characteristic curve of 92%. This fully automated method allows for sensitive and specific detection of lesion incidence that can be applied to large collections of images.

email: emsweene@jhsph.edu

5l. COMPARING INDEPENDENT COMPONENT ANALYSIS ESTIMATION METHODS WITH AN APPLICATION TO NEUROIMAGING OF RESTING STATE FUNCTIONAL CONNECTIVITY IN ATTENTION DEFICIT AND HYPERACTIVITY DISORDER

Benjamin B. Risk, Cornell University*
David S. Matteson, Cornell University
David Ruppert, Cornell University

Independent component analysis (ICA) is a common approach to identify independent sources from multivariate observations. Applications include feature extraction in brain imaging data in order to characterize the pathophysiology of neurological disorders. Independent components can be estimated by maximizing distance from normality (fastICA), minimizing fourth-order cross cumulants (JADE), combining density estimation with profile MLE (ProDenICA), and other techniques. A recent approach by Matteson and Tsay uses a novel contrast function based on the distance covariance between components (dCovICA). FastICA and JADE contrast functions are based on necessary but not sufficient conditions for independence and are only consistent for certain symmetric distributions. Both ProDenICA and dCovICA are based on necessary and sufficient conditions, where ProDenICA has smoothness constraints, and dCovICA is consistent for all distributions with finite second moments. Consistency and sufficient conditions for independence are particularly important in applications such as modeling resting state fMRI, in which the source distributions are unknown. We first examine the performance of ICA methods via simulation studies with known source distributions. Then we apply these methods to fMRI data to examine differences between healthy and ADHD subjects.

email: bbr28@cornell.edu

5m. VARIABLE SELECTION METHODS IN LINEAR MODELS WITH GROWING DIMENSION

June Luo, Clemson University*

High dimensional data with dimension p far bigger than the sample size n has received extensive attention with the proliferation of microarray data. Objectives include parameter estimation and variable selection. In this presentation, theoretical approach for variable selection is of interest. I will start with a linear regression, apply the ridge estimation method, derive the asymptotic distribution of the ridge estimator and propose consistent screening procedures under different scenarios. The asymptotic properties are obtained as dimension going to infinity and sample size being fixed. A new test statistic for simultaneous high dimension coefficients test is developed. The established theory is extended to partial linear models after I adopt the differencing techniques.

email: jluo@clemson.edu

5n. FAMILY-BASED ASSOCIATION STUDIES FOR NEXT-GENERATION SEQUENCING

Yun Zhu, University of Texas, Health Science Center at Houston*
Momiao Xiong, University of Texas, Health Science Center at Houston

Fast and cheaper next generation sequencing (NGS) technologies will generate high-dimensional genetic variation data that allow nearly complete evaluation of genetic variation. Despite their promise, NGS technologies also suffer from three remarkable limitations: high error rates, enrichment of rare variants and large proportion of missing values. To meet analytic challenge raised by NGS and new disease models, we propose here a general framework for sequence-based association studies which can use either population or family or both population and family data sampled from any population structure. We also propose a universe procedure that can transform any population-based association test statistics to family-based association tests and develop family-based functional principal component analysis (FPCA) with or without smoothing, generalized, CMC and single marker association test statistics. By simulations, we demonstrate that the family-based smoothed FPCA has much higher power to detect association than other population-based or family-based association analysis methods. To further evaluate its performance, the proposed statistics are applied to GWAS dataset with only rare variants in Framingham Heart Study with pedigree structures.

email: winonone@gmail.com

5o. CASE-BASED REASONING IN COMPARATIVE EFFECTIVENESS RESEARCH

Marianthi Markatou, T. J. Watson Research Center, IBM*
Prabhani Kuruppumullage Don, The Pennsylvania State University

Given a new patient with a specific medical diagnosis can we identify other, similar to the new, patients in order to inform the initiation of care in this individual? We interface methods from statistics and artificial intelligence to address the aforementioned question. Specifically, we discuss the applicability of the case-based reasoning framework, originated in the cognitive sciences, and we develop its statistical foundations appropriate for comparative effectiveness research. Key issues that we address in the development of the statistical underpinnings of case-based reasoning are the concepts of patient representation and patient similarity. A variety of statistical distances are used as similarity measures and we discuss their bias-variance decompositions. We present an algorithm, based on statistical distances, that identifies similar to the new subjects, and we discuss the trade-off between statistical power and degree of similarity. This algorithm imitates the key strength of the randomized control trials in that, it balances the distribution of the covariates, and performs well with high dimensional data. Simulations and an example comparing the effectiveness of statins for the treatment of hyperlipidemia, further illustrate its performance in terms of controlling bias and confounding.

email: mmarkat@us.ibm.com

5p. REGULARIZATION WITH LATENT FACTORS FOR MODEL SELECTION IN MULTIVARIATE MULTIPLE REGRESSION WITH APPLICATION TO eQTL ANALYSIS

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

We propose a novel method to search important predictors in multivariate multiple linear regressions with latent factors. This study is mainly motivated by Genome-Wide Association Study (GWAS), which is to identify specific SNPs that affect gene expressions. Generally, GWAS is greatly challenged by many complicating factors, including (i) multi-collinearity (or redundant linkages) among SNPs in linkage disequilibrium, (ii) complicated SNP-gene expression relationships, (iii) a very large number of SNPs, (iv) a small sample size, and (v) unobserved non-genetic factors, such as population structures, or environmental conditions. Moreover, adjusting for non-genetic factors will allow us to gain statistical power in the analysis. Thus, it is critical to apply appropriate tools to extract signals in such massive data. In this study, we model the gene-SNP associations through multivariate multiple regressions with latent factors. In addition, we develop a group-wise coordinate descent algorithm to solve the related optimization problem efficiently. Criteria for selecting tuning parameters are also discussed. By extensive simulation studies, it shows that our method can achieve high sensitivity and specificity, while, can improve the statistical power to detect important predictors. Finally, data analysis is included for illustration.

email: zhouyan@umich.edu

5q. MEASUREMENT ERROR MODEL IN SHAPE ANALYSIS

Jiejun Du, University of South Carolina*
Ian Dryden, University of South Carolina
Xianzheng Huang, University of South Carolina

In statistical shape analysis, the original shape data are often prone to measurement error; this is conventionally ignored in shape analysis when we calculate a non-Euclidean distance between two shapes or match one object to another. In this study, we consider the measurement error models for shape data (two or three dimensional data). First we show that the naive ordinary least squares estimator is biased when the measurement error is ignored. Then we consider structural measurement error models for shape data and derive improved estimators for the parameters in the model. The methodology is illustrated with an application in forensic face identification.

email: jiejun.du@gmail.com

6. ENVIRONMENTAL, EPIDEMIOLOGICAL, HEALTH SERVICES AND OBSERVATIONAL STUDIES

6a. PUBLICATION BIAS IN META-ANALYSIS

Min Chen, ExxonMobil Biomedical Sciences, Inc.*

Meta-analysis is the statistical synthesis of results of several studies that address a set of related research hypotheses. The synthesis will be meaningful only if the studies have been obtained systematically. However, publication bias can be an issue for meta-analysis. Studies with significant results are more likely to be published than those with non-significant results, and published studies are more likely to be included in a meta-analysis than unpublished studies (Borenstein et al. 2009). In this research, I address ways to assess and correct for publication bias in a benzene/leukemia meta-analysis (Khalade et al. 2010) using various statistical methods and approach described in Rothstein et al. (2005), including the funnel plot, Begg and Mazumdar rank correlation test, Egger's test of the intercept, Rosenthal's Fail-safe N, Orwin's Fail-safe N, and Duval and Tweedie's Trim and Fill.

email: min.chen@exxonmobil.com

6b. REGRESSION MODELS FOR GROUP TESTING DATA WITH POOL DILUTION EFFECTS

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

Group testing, since its advent, has been widely used to reduce the cost associated with infectious disease screening. Although, group testing is an intuitive method of reducing the time and cost associated with screening large populations for rare infections, it has proven to be an invaluable tool in estimating the proportion of infected individuals in a population. In the more recent literature, regression models for pooled response data have been proposed as a way to incorporate covariates in the estimation of the infection prevalence. These regression methods, like many other group testing estimation techniques, proceed under the assumption that testing error rates for pooled specimens are the same as those for individual specimens. This assumption is a topic of key concern among the proponents of pooled testing, because it is natural to believe that the testing accuracies for a pooled specimen may depend on the number of specimens included in the pool. Subsequently, the focus of this research is to develop a pooled testing regression model that incorporates pool specific testing error rates into the estimation process. To assess the characteristics of our new methodology we apply our regression model to Hepatitis B data collected from prisoners in Ireland.

email: mcmahanc@mailbox.sc.edu

6c. PRINCIPAL STRATIFICATION BASED ON LATENT SURVIVAL CLASSES TO PREDICT TREATMENT OUTCOMES FOR LOCALIZED KIDNEY CANCER

Brian L. Egleston, Fox Chase Cancer Center*
Yu-Ning Wong, Fox Chase Cancer Center
Robert G. Uzzo, Fox Chase Cancer Center

Rates of kidney cancer have been increasing, with localized small incidental tumors experiencing the fastest growth rates. Much of the increase could be due to the increased use of CT scans, MRIs, and ultrasounds for conditions unrelated to kidney cancer. Many of these tumors might never have been detected or become symptomatic in the past. This suggests that many patients might benefit from less aggressive therapy, such as partial rather than total removal of the kidney. It is even possible that some patients do not need treatment for localized kidney cancer. In this work, we propose using a principal stratification framework to estimate the proportion and characteristics of individuals who have large or small hazard rates of death in two treatment arms. This will allow us to predict who might be helped or harmed by aggressive treatment. We will use Weibull mixture models as part of the project. This work differs from much previous work in that the survival classes upon which principal stratification is based are latent variables. That is, survival class is not a variable observed in the data. We will apply this work to estimate surgical treatment effects using linked Surveillance Epidemiology and End Results-Medicare claims data.

email: Brian.Egleston@fccc.edu

6d. LOSS FUNCTIONS FOR IDENTIFYING REGIONS WITH MINIMUM OR MAXIMUM RATES

Ronald E. Gangnon, University of Wisconsin-Madison*

Excess mortality in a population of interest is the number of deaths that should not have occurred or could have been avoided under different conditions. In application, excess mortality is assessed relative to the observed mortality rate in some well-performing subgroup of the population, e.g. the best county in the United States. Thus, correct identification of the best performing subgroup is of particular interest. We construct several loss functions that address this goal and evaluate candidate estimates of the region with the minimum rate based on these loss functions. We compare inferences using the candidate estimators using data from the County Health Rankings project. We discuss extensions of these loss functions to more general ranking problems.

email: ronald@biostat.wisc.edu

6e. TIME-TO-EVENT ANALYSIS OF AMBIENT AIR POLLUTION AND PRETERM BIRTH

Howard H. Chang, Emory University*
Brian J. Reich, North Carolina State University
Marie Lynn Miranda, University of Michigan

Preterm birth is associated with significant neonatal morbidity and mortality, in addition to long term health and developmental complications. We describe a spatial discrete-time survival model to estimate the effect of air pollution on the risk of preterm birth. The standard approach treats prematurity as a binary outcome and cannot effectively examine time-varying exposures during pregnancy. We view gestational age as time-to-event data where each pregnancy enters the risk set at a pre-specified time (e.g. the 27th week). The pregnancy is then followed until either (1) a birth occurs before the 37th week (preterm); or (2) it reaches the 37th week and a full-term birth is expected. The survival approach allows us to examine both long-term and short-term effects of air pollution using cumulative and lagged exposure metrics. We applied the survival model to examine the association between exposure to fine particle air pollution during pregnancy and the risk of preterm birth in North Carolina between the periods 2001 to 2005.

email: howard.chang@emory.edu

6f. DORFMAN GROUP SCREENING WITH MULTIPLE INFECTIONS

Yanlei Peng, University of South Carolina*
Joshua M. Tebbs, University of South Carolina
Christopher R. Bilder, University of Nebraska-Lincoln

Group (pooled) testing is commonly used to reduce the cost associated with screening individuals for infectious diseases. A great deal of work in the literature addresses group testing classification in the context of a single infection, including methods to obtain optimal pool sizes and to evaluate testing accuracy. However, there are currently no methods available when screening for multiple diseases simultaneously. Subsequently, the focus of our work is to extend Dorfman's methods to multiple infections. More explicitly, we obtain the operating characteristics of Dorfman's decoding procedure when an assay can test simultaneously for two diseases. We are able to determine the optimal pool size that minimizes the expected number of tests. Moreover, we show that significant gains in testing efficiency can be realized if the two infection statuses are highly correlated. Additionally, with respect to different infections, we are able to obtain closed-form expressions for testing error rates. Our work is motivated by the screening practices at the University of Iowa Hygienic Laboratory, where nucleic acid amplified technology is used to screen individuals simultaneously for chlamydia and gonorrhea.

email: peng@email.sc.edu

6g. APPLYING GENERAL RISK SCORES IN SPECIAL POPULATIONS

Cynthia S. Crowson, Mayo Clinic*
Elizabeth J. Atkinson, Mayo Clinic
Terry M. Therneau, Mayo Clinic

Risk prediction tools are increasingly being used to estimate individuals' risks of developing disease in the clinical setting. Often these risk scores are developed to predict the risk in the general population (e.g., Framingham study), but they may not accurately assess risk in patients with particular comorbidities (e.g., rheumatoid arthritis). We show how to assess the accuracy (i.e., calibration-in-the-large) of these general risk scores in special populations, as well as how to use the general risk scores to assess whether special patient populations have a higher risk of disease than would be expected for persons with the same risk factor profile in the general population. The basic approach is to transform the predicted risks to expected numbers of events, which can then be manipulated using the array of methods that have been developed for adjusted survival. This approach augments the capabilities of the methods commonly used to assess calibration of risk scores. Using this approach we can often avoid the need to collect per-study controls.

email: crowson@mayo.edu

6h. USING PREDICTIVE SURFACES TO UNDERSTAND DISPARITIES IN EXPOSURE TO PM2.5 AND OZONE IN NORTH CAROLINA

Simone Gray, U.S. Environmental Protection Agency*
Sharon Edwards, University of Michigan
Marie Lynn Miranda, University of Michigan

A large body of research has shown that disparities exist not only in the health status between different low-income groups and racial minorities but also in the distribution of environmental risks and hazards. To investigate the relationship between air pollution (AP) exposure and human health, AP measurements from monitoring stations are commonly used as a proxy for personal exposure. In this study, we use modeled predictive surfaces to examine the relationship between AP exposure and measures of socio-economic status in North Carolina. The daily measurements of PM2.5 and O3 are calculated through a spatial Hierarchical Bayesian model that uses data from two sources as inputs: monitoring networks and numerical models. The numerical models produced gridded AP concentrations while the monitoring data is measured at a fixed point source. The "fused" model combines the two data sources to produce predictions for 12km domains across the entire state of North Carolina. Using the fused data as a surrogate for AP exposure, we look closely at the demographics of communities in areas with higher levels of environmental hazards, paying particular attention to the associated measures of SES and racial composition of these areas.

email: simonecgray@gmail.com

6i. PROBABILISTIC RISK ASSESSMENT OF AIR QUALITY MANAGEMENT STRATEGIES FOR OZONE

Kristen M. Foley, U.S. Environmental Protection Agency*
Brian J. Reich, North Carolina State University
Sergey L. Napelenok, U.S. Environmental Protection Agency

The US EPA uses numerical air quality models to design emission control strategies for improving ambient ozone concentrations across the US. A combination of deterministic air quality models and Bayesian statistical methods are used to derive probabilistic estimates of air quality. The statistical framework quantitatively incorporates uncertainty information about important inputs to the numerical air quality model. The probabilistic model predictions are weighted based on population density in order to better quantify the societal benefits/disbenefits of different air quality management strategies. Four different emission control strategies are compared. The probabilistic approach provides more information for cost-benefit analysis and regulatory impact analysis compared to standard methods that do not account for the variability and uncertainty in the atmospheric systems being modeled. The statistical model is shown to be well calibrated compared to observed ozone levels using cross validation.

email: foley.kristen@epa.gov

6j. A SIMULATION STUDY OF ESTIMATORS OF TIME-VARYING TREATMENT EFFECTS ON CANCER RECURRENCE WITH TIME DEPENDENT CONFOUNDING

Jincheng Shen, University of Michigan*
Edward H. Kennedy, VA Center for Clinical Management Research
Douglas E. Schaubel, University of Michigan
Lu Wang, University of Michigan
Jeremy M.G. Taylor, University of Michigan

Marginal structural models (MSM) provide a powerful tool to estimate the causal effect of a treatment by appropriately controlling for time-dependent confounding (Hernán et al., 2000). In a prostate cancer dataset, where the treatment is time-varying, and its effect is confounded by the patient's time trajectory of prostate-specific antigen (PSA) levels, methods have been proposed to estimate the subject-specific treatment effect (Edwards et al., 2010). We conduct several simulation studies along with this prostate cancer setting to assess the properties of the treatment effect estimated by the MSM method and investigate how it is related to the subject-specific treatment effect. Furthermore, we consider several randomized trial scenarios and link this marginal treatment effect to treatment effects that would be of scientific interest.

email: jcshen@umich.edu

6k. PROPENSITY SCORE USING MACHINE LEARNING

Yi-Fan Chen, University of Pittsburgh*
Lisa Weissfeld, University of Pittsburgh

The propensity score (PS) is often used for analyses in observational studies. It provides a tool for creating similar groups within a cohort that are later used for comparative analyses. Recently, due to the interest in comparative effectiveness research (CER), there is interest in understanding the performance of the propensity score when comparing potential treatment groups. Propensity scores can be created either through the use of logistic regression, or through machine learning techniques. In recent years, machine learning techniques have been shown to result in gains in efficiency and a reduction in bias, in addition to relaxing parametric assumption. However, these methods have not been rigorously studied over a wide range of potential scenarios. In this study, we evaluate different PS estimation approaches in various settings via simulations. We focus on the false-positive rate when no treatment effect exists. We demonstrate how false findings can occur if PS is inappropriately used and recommend reasonable approaches under different scenarios.

email: yic33@pitt.edu

6l. ASSESSING THE EFFECT OF ORGAN TRANSPLANT ON THE DISTRIBUTION OF RESIDUAL LIFETIME

David M. Vock, North Carolina State University*
Anastasios A. Tsiatis, North Carolina State University
Marie Davidian, North Carolina State University

Because of the scarce number of organs available for transplantation in the United States, it is important to quantify the expected change in residual lifetime as a result of transplantation to improve organ allocation. However, there has been little work to assess the survival benefit of transplantation from a causal perspective. Across all organ transplants, patients with worse waitlist prognosis are given preference when an organ becomes available; the time-varying covariates that affect waitlist mortality and transplant likelihood must be accounted for to obtain valid causal inference of survival benefit. Previous methods developed to estimate the causal effects of treatment in the presence of time-varying confounders have assumed that treatment assignment was independent across patients, which is not true for organ transplantation. We develop a version of G-estimation that accounts for the fact that treatment assignment is not independent across individuals to estimate the parameters of a structural nested failure time model. We demonstrate our method on the survival benefit of lung transplantation.

email: dmvock@ncsu.edu

6m. METHODS FOR CLASSIFYING CHANGES IN BACTERIAL PREVALENCE OVER TIME

Raymond G. Hoffmann, Medical College of Wisconsin*
Ke Yan, Medical College of Wisconsin
Pippa Simpson, Medical College of Wisconsin
Jessica Vandevalle, University of Wisconsin - Milwaukee
Sandra McLellan, University of Wisconsin - Milwaukee

Information about the prevalence of bacterial taxa in water samples can be determined via next generation sequencing. Samples were taken from two different, but related, sites in Lake Michigan over a three year period. After aggregation of similar taxa there were 22 time series of bacterial prevalence x two sites. The goal was to identify the taxa that had similar temporal patterns. Wavelet analysis after filtering to reduce sampling noise was used to determine the temporal aspects of the multivariate time series because of the irregular time measurements and the shape of the prevalence curve. Comparisons of clustering of the wavelet coefficients, the coefficients of harmonic regression and classification tree (CART) with Random Forests are made to identify stability and utility of the classifications.

email: rhoffmann@mcw.edu

6n. SALIVARY CORTISOL AS A PREDICTOR OF HEALTH OUTCOMES

Brisa N. Sanchez, University of Michigan*
Ana V. Diez-Roux, University of Michigan
TE Raghunathan, University of Michigan

Epidemiological studies now frequently collect salivary cortisol measures as an objective measure of stress to better characterize the contribution of stress to disease and the mediating role of stress in race/ethnic or socio-economic health disparities. Salivary cortisol exhibits a nonlinear pattern throughout the length of the day, which, given repeated measures within subjects, can be modeled as a functional response using, for example, functional mixed models. When relating stress as a predictor of disease, linear models with functional predictors can be used. However, comparing results from these types of models to existing literature is difficult, since existing literature is based on summary measures of the salivary response curve, such as the area under the curve. We describe approaches by which inferences on commonly used measures of the curve can be derived directly from models using the response curve as a functional predictor. We compare this approach to simple two-step approaches in terms of bias, efficiency and MSE, but also in terms of practical implementation, and demonstrate the use in real data examples.

email: brisa@umich.edu

6o. RETRACING MICRO-EPIDEMICS OF CHAGAS DISEASE USING EPICENTER REGRESSION

Michael Levy, University of Pennsylvania
Dylan Small, University of Pennsylvania*
Joshua Plotkin, University of Pennsylvania

Vector-borne transmission of Chagas disease has become an urban problem in Arequipa, Peru, yet the debilitating symptoms that can occur in the chronic disease are rarely seen in hospitals in the city. The lack of clinical disease in Arequipa has led to speculation that the local strain of the etiologic agent, Trypanosoma cruzi, has low chronic pathogenicity. The long asymptomatic period of Chagas disease leads us to an alternative hypothesis for the absence of clinical cases in Arequipa: transmission in the city may be so recent that most infected individuals have yet to progress to late stage disease. Here we describe a new method, epicenter regression, which allows us to infer the spatial and temporal history of disease transmission from a snapshot of a population's infection status. We show that in a community of Arequipa, transmission of Trypanosoma cruzi occurred as a series of focal micro-epidemics. Most extant human infections in this community arose over a brief period of time immediately prior to vector control. According to our findings, the symptoms of chronic Chagas disease are expected to be absent, even if the strain is pathogenic, given the long asymptomatic period of the disease and short history of intense transmission.

email: dsmall@wharton.upenn.edu

6p. COMPARING CANCER RATES BY AGE-STRATIFIED ZERO-INFLATED POISSON MODEL

Xiaoqin Xiong, Information Management Services, Inc.*
Binbing Yu, National Institute on Aging, National Institutes of Health

The annual percent change (APC) has been used as a measure to describe the trend in the age-adjusted cancer incidence or mortality rate over relatively short time intervals. To compare the recent cancer trends by gender or by geographic regions, one compares their APCs. The traditional method to estimate the APC was by fitting an ordinary or weighted linear regression model of the logarithm of age-adjusted rates. To compare the APCs of two regions, two slopes were compared by t-test or modified t-test. Another method was by modeling the mortality or incidence counts using a stratified Poisson model. Then a corrected Z-test was applied to test the equality of two APCs. Recently, the cancer incidence rates for rare cancers and for small geographical areas are of interest to local policy maker and health researchers. For these data, there are an excessive number of zeros in incidence counts. The use of Zero-inflated Poisson (ZIP) model is more favorable than Poisson model. We propose an age-stratified ZIP model for comparing trends in cancer rates across two overlapping regions. The proposed method is illustrated using age-adjusted cancer incidence data from the Surveillance, Epidemiology and End Results (SEER) Program of the National Cancer Institute.

email: jenemyxiong@yahoo.com

6q. USE OF THE CONTINUOUS-TIME MARKOV CHAIN TO EXAMINE THE NATURAL HISTORY OF ALZHEIMER'S DISEASE

Wenyaw Chan, University of Texas, Health Science Center at Houston*
Julia Benoit, University of Texas, Health Science Center at Houston
Rachelle S. Doody, Baylor College of Medicine

Alzheimer's disease (AD) is one of the leading causes of death among U.S. adults aged 65 and older. Current research into determinants of AD progression is often based on longitudinal regression models or longitudinal generalized linear models. Although these methods can elucidate the relationship between covariates and cognitive or functional measures, they cannot be extended to describe the nature process of disease progression. In this study, we propose a continuous-time Markov chain model approach that describes future changes in the disease stage (mild, moderate or severe) conditioned on the current stage. The Markov model will also be applied to identify covariates associated with the nature process of disease progression. An explicit likelihood function will be presented and an iterative computational algorithm will be discussed. The proposed methods will be applied to data collected in a longitudinal cohort study of Alzheimer's disease at the Alzheimer's Disease and Memory Disorders Center at Baylor College of Medicine.

email: wenyaw.chan@uth.tmc.edu

6r. PROCESS-BASED BAYESIAN MELDING OF TWO-ZONE MODELS AND INDUSTRIAL WORKPLACE DATA

Joao V.D. Monteiro, University of Minnesota*
Sudipto Banerjee, University of Minnesota
Gurumurthy Ramachandran, University of Minnesota

A primary issue in industrial hygiene is the estimation of a worker's exposure to chemical, physical and biological agents. Mathematical models, based upon physical principles, are being increasingly used as a method for assessing occupational exposures. However, predicting exposure in real workplace settings is constrained by lack of quantitative knowledge of exposure determinants. Non-linear regression models for the two-zone differential equation models are less effective for predictive inference as they do not account for biases in the physical model attributable, at least in part, to extraneous variability. We recognize these limitations and provide a rich and computationally effective Bayesian hierarchical framework that melds the physical model with the observed workplace 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 formulate two different approaches: (i) where the inputs to the physical model are unknown and must be estimated, and (ii) where the inputs are controlled and we seek to meld the output from the physical model with the observed data.

email: monte092@umn.edu

7. CORRELATED AND LONGITUDINAL DATA

7a. MULTILEVEL JOINT ANALYSIS OF LONGITUDINAL AND BINARY OUTCOME

Seo Yeon Hong, University of Pittsburgh*
Lisa A. Weissfeld, University of Pittsburgh

Joint modeling has become a topic of great interest in recent years. The focus has been primarily on models that include a longitudinal and a survival component. The models are simultaneously analyzed using a shared random effect that is common across the two components. While these methods are useful when time-to-event data are available, there are many cases where the outcome of interest is binary and a logistic regression model is used. This work focuses on the setting where the longitudinal outcome is subject to multiple hierarchical levels and the second outcome is binary. We propose the use of a joint model with a logistic regression model being used for the binary outcome and a hierarchical mixed effects model being used for the longitudinal outcome. This model is applicable in setting where data are hierarchically structured, such as longitudinal measures nested within subject and then nested within hospital. We link the two sub-models using both subject and cluster level random effects and compare it with models using only one level of random effects. We use the Gaussian quadrature technique implemented in a ML (Multiprocess Multilevel Modeling software). Simulation studies are presented to illustrate the properties of the proposed model.

email: seh72@pitt.edu

7b. SPACE-TIME HETEROGENEITIES IN ONE-DIMENSIONAL POINT PROCESS DATA: MODELING SEA TURTLE NESTING PATTERNS VIA LOG-GAUSSIAN COX PROCESSES

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

Spatial-temporal point pattern data are increasingly available, including one-dimensional, directionless observations observed along a line. Our motivating example involves sea turtle nesting data with space and time-specific emergence locations along Juno Beach, Palm Beach County, Florida for the years 1998-2000. Our analytic goal is to assess spatial and temporal heterogeneities in the local probabilities of emergence under the framework of Log-Gaussian Cox Processes (LGCPs). The non-parametric estimates of the intensity and the second-order characteristics, i.e., K-function and pair correlation function are investigated to exhibit spatial and temporal clustering point patterns and their evolution over the course of individual nesting seasons and as impacted by local beach construction projects. We also consider parametric modeling of LGCPs via minimum contrast method and critically assess the assumption of space and time separability. Finally, we review model checking diagnostics to examine the appropriateness of our estimated LGCP models.

email: mwang36@emory.edu

7c. JOINT MODELING OF LONGITUDINAL MULTIVARIATE MEASUREMENTS AND SURVIVAL DATA WITH APPLICATIONS TO PARKINSON'S DISEASE

Sheng Luo, University of Texas at Houston*
Bo He, University of Texas at Houston

In large-scale clinical trials on chronic Parkinson's disease, longitudinal multivariate measurements of mixed types are often collected to monitor patients' disease status and assess the global treatment effect. However non-ignorable missing not at random (MNAR) occurs in longitudinal measurements due to some survival events (e.g., death or time to symptomatic treatment). The longitudinal outcomes can be subject to measurement errors. We propose a joint modeling framework for analysis of both longitudinal multivariate outcomes and survival outcome to evaluate the treatment effect on both longitudinal and survival

outcomes, simultaneously and the effect of longitudinal outcomes on survival time. An item response model with latent variables indicating patient's disease severity is applied to account for all sources of correlation in longitudinal outcomes. The Cox's model with piecewise constant baseline hazard function is used for survival outcomes. Inference is conducted using a Bayesian framework via Markov chain Monte Carlo simulation. This methodological development has been motivated and applied to a major Parkinson's disease trial--DATATOP, where we are interested in the effect of deprenyl on both the patients' disability outcomes and the time to levodopa therapy.

email: Bo.He@uth.tmc.edu

7d. IDENTIFICATION OF CLINICALLY RELEVANT DISEASE SUBTYPES USING SUPERVISED SPARSE CLUSTERING

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

Conventional clustering algorithms often produce poor results when applied to high-dimensional data sets where many of the predictor variables are unrelated to the clusters. Several recent studies have shown that the accuracy of clustering methods can be improved by clustering based on a subset of the predictors or by assigning an appropriate weight to each predictor. However, in many applications, one wishes to identify clusters that are associated with an outcome of interest, and this problem has not been studied extensively. We propose a method for identifying clusters associated with an outcome variable in high-dimensional data using a modified version of the "sparse clustering" method of Witten and Tibshirani (2010), which is a weighted version of k-means clustering. Although this method performs well in many situations, it is not guaranteed to identify clusters that are associated with an outcome variable. We modify this method by giving greater initial weights to predictors that are most strongly associated with the outcome variable. We show that our proposed method outperforms several competing methods on a series of simulated data sets. We also show how our method can be used to identify clinically relevant clusters in a cohort of patients with chronic pain.

email: smgaynor@live.unc.edu

7e. MULTISTATE MARKOV CHAIN TRANSITION MODEL FOR CLUSTERED LONGITUDINAL DATA: APPLICATION TO AN OSTEOARTHRITIS STUDY

Ke Wang, Boston University*
Bin Zhang, Boston University
Yuqing Zhang, Boston University
Haiqun Lin, Yale University
Howard Cabral, Boston University

Multistate Markov chain models are useful tools for analyzing fluctuation of categorical disease outcomes in longitudinal studies. However, in some musculoskeletal research, outcomes are measured on multiple joints of each individual. Joints of a same person form a cluster and within-cluster correlation needs to be accounted for. Here we propose multistate Markov chain transition models that take within-cluster correlation into consideration. Regression coefficients are estimated by maximizing the likelihood and we use three marginal-model-based approaches to account for within-cluster correlation. The first method is based on a clustered approach that utilize a GEE robust sandwich estimator, the second method is a within-cluster resampling approach that randomly select one knee from each cluster, and the third method is based on a cluster-weighted estimation equation. Simulated data sets are used to evaluate the performance of the approaches. Three scenarios were simulated, corresponding to data with weak correlation, with strong correlation, and with cluster size dependent on risk of the outcome. We then used the models to analyze knee pain severity data from a longitudinal knee osteoarthritis study in the US (The Osteoarthritis Initiative).

email: kewang@bu.edu

7f. A BIVARIATE LOCATION-SCALE MIXED-EFFECTS MODEL WITH APPLICATION TO ECOLOGICAL MOMENTARY ASSESSMENT (EMA) DATA

Oksana Pugach, University of Illinois at Chicago*
Donald Hedeker, University of Illinois at Chicago

EMA yields longitudinal data with many measurements per subject, and often with several outcomes measured simultaneously. Such data can be modeled by specifying a joint mixed-effects model with relaxed assumptions on the homogeneity of within-subject (BS) and between-subject (WS) variances. In this presentation, two continuous measurements of mood are modeled using a bivariate mixed-effects linear model. The variance-covariance matrices of the BS and WS effects are modeled in terms of covariates, explaining BS and WS heterogeneity. Furthermore, the WS variance models are extended by including random scale effects. For illustration, data from a natural history study on adolescent smoking are used in analysis. 431 students, from 9th and 10th grades, reported on their mood and activities at random prompts

during seven consecutive days. This resulted in 14,105 prompts (average of 30 per student). Positive Affect (PA) and Tired/Bored (TB) measures were two outcomes modeled jointly. Results of the analyses suggest that female students had larger variation in their PA mood than male student, whereas the WS variance in the tired/bored measure did not exhibit significant gender differences. The WS and BS covariance for the two outcomes were negative and significant.

email: opugac1@uic.edu

7g. THE INFLUENCES OF UTILIZED AND THEORETICAL COVARIANCE WEIGHTING MATRICES ON THE ESTIMATION PERFORMANCE OF QIF

Philip M. Westgate, University of Kentucky*

The method of Quadratic Inference Functions (QIF) is increasingly popular for the marginal analysis of correlated data due to its multiple advantages over generalized estimating equations. Asymptotic theory is used to derive analytical results from QIF, and therefore it is important that the finite-sample properties of this method are understood. In particular, we give focus to the variances of final parameter estimates. Three asymptotically equivalent weighting matrices are seen in the QIF literature, and we consider their finite-sample differences with respect to QIF's estimation performance. This then leads to an examination of the importance of using accurate initial parameter estimates inside the empirical covariance matrix that is utilized in practice with QIF's estimating equations, and whether or not these should be continuously updated during the iterative estimation procedure. These issues are demonstrated via an applied dataset and simulation in general repeated measures and cluster randomized trials settings.

email: philip.westgate@uky.edu

7h. GROUP-BASED TRAJECTORY MODELING OF CARDIAC AUTONOMIC MODULATION

Michele Shaffer, Penn State College of Medicine*
Fan He, Penn State College of Medicine
Duanping Liao, Penn State College of Medicine

Existing literature supports a circadian pattern of cardiac autonomic modulation (CAM). Heart rate variability (HRV) is regulated by the balance of sympathetic and parasympathetic modulations, and it is a commonly used noninvasive measurement of CAM. Using the Air Pollution and Cardiac Risk and its time course (APACR) study, we investigated the HRV trajectories of 101 individuals using group-based trajectory modeling with a normal distribution for continuous outcomes and a Poisson distribution for count outcomes. Group-based trajectory modeling is an application of finite mixture models to identify clusters or subgroups of individuals following similar patterns over time. Previous investigations of this data utilized cosine periodic regression, which can be summarized with three parameters: the mean, amplitude, and acrophase and compared groups with specific covariate patterns. In contrast, group-based trajectory modeling assumes the average value changes smoothly over time and a priori the number of potential subgroups was not identified. Models investigated included the time-independent variables age at baseline, sex, ethnicity, and cardiovascular disease-related conditions. Findings are compared between the periodic regression and group-based trajectory models.

email: shaffer.michele@psu.edu

7i. STATISTICAL INFERENCE ON TEMPORAL GRADIENTS IN REGIONALLY AGGREGATED DATA

Harrison S. Quick, University of Minnesota*
Sudipto Banerjee, 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. Our major objective is to carry out inference on gradients of the temporal process, while at the same time accounting for spatial similarities of the temporal process across neighboring regions. 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 cross-covariance function we can ensure that the temporal process realizations are mean square differentiable, and may thus subsequently carry out inference on temporal gradients in a posterior predictive fashion. We use this approach to evaluate temporal gradients in a dataset comprised of monthly county level asthma hospitalization rates in the state of California 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: quic0038@umn.edu

7j. MODELING AND ESTIMATION OF REPEATED ORDINAL DATA USING GAUSSIAN COPULA

Raghavendra R. Kurada, Old Dominion University
Roy T. Sabo, Virginia Commonwealth University
N. Rao Chaganty, Old Dominion University*

Epidemiological and medical studies often involve repeated measurements on independent subjects, for which a wide variety of statistical methodologies are available. However, theoretical and practical challenges arise for repeated-measure categorical data, especially when the outcomes are ordinal in nature. In this talk we will discuss a latent variable likelihood methodology for repeated-measure ordinal data using Gaussian copula with probit and logit link functions. We derive the score functions and simplified expressions for the Hessian matrices, which allow easy computation of the standard errors for the marginal regression parameter estimates as well as the dependence parameters. Through asymptotic relative efficiency calculations we demonstrate that these likelihood estimators are superior as compared to estimators arising from previously established estimating equation approaches. We apply this likelihood-based methodology in an analysis of two real-life data examples using an R package developed specifically for the likelihood estimation.

email: rchagant@odu.edu

7k. SIMULATION STUDY OF THE CONVERGENCE PROPERTIES OF LOG GAUSSIAN COX PROCESS POSTERIORS

Timothy D. Johnson, University of Michigan
Ming Teng, University of Michigan*
Jian Kang, Emory University

Log Gaussian Cox Processes (LGCP) has proved to be quite useful in fitting spatial point pattern data, due in large part to their mathematical tractability. Moller, et al. (1998) presented an algorithm to estimate the underlying intensity function within the Bayesian framework. Later (2005), Waagepetersen proved that the approximate posterior (by discretizing the Gaussian process) converges in expectation to the true posterior distribution as the cells, which make up the discretized grid, tend uniformly to zero. He showed by way of example that the posterior mean and variance appear to be sensitive to grid size. To date, we are unaware of any systematic study of this sensitivity. Thus, we conducted a simulation study to determine the relationship between the grid size and RMSE of the various parameters in the model. We also adopt a frequentist view for our simulation study and set parameters at fixed values and study the bias of our posterior means as well as coverage of posterior credible intervals. We show that if the marginal variance of the LGCP is too small, then the correlation parameter cannot be recovered. We demonstrate our results on real data and give guidelines regarding grid size.

email: tengming@umich.edu

7l. SENSITIVITY OF A LONGITUDINAL ANALYSIS TO MISSING DATA HYPOTHESES: A STUDY OF THE MECHANISMS BY WHICH WEIGHT LOSS REDUCES ARTERIAL STIFFNESS

Jennifer N. Cooper, University of Pittsburgh*
Jeanine M. Buchanich, University of Pittsburgh
Ada Youk, University of Pittsburgh
Maria M. Brooks, University of Pittsburgh
Kim Sutton-Tyrrell, University of Pittsburgh

Arterial stiffness decreases with weight loss in overweight and obese adults. We aimed to determine the mechanisms by which this occurs. Because follow-up data are not likely to be missing at random in lifestyle intervention studies, we used pattern-mixture modeling and Markov Chain Monte Carlo multiple imputation to evaluate the influence of different missing data assumptions on our findings. Repeated measures of arterial stiffness were collected in overweight/obese young adults participating in a behavioral weight loss intervention. At the 6 and 12 month follow-up time points, 17% and 26% of participants were missing arterial stiffness data. Linear mixed effects models were used to evaluate associations between weight loss and arterial stiffness and to examine the degree to which improvements in obesity-related factors explained these associations. Weight loss appeared to improve aortic and peripheral arterial stiffness in overweight/obese young adults by different mechanisms. Missing data, though hypothesized to be non-ignorable, did not appear to substantially impact the findings obtained from mixed models.

email: jnn9@pitt.edu

7m. MULTIVARIATE SPATIAL ANALYSIS VIA MIXTURES

Brian Neelon, Duke University*
Rebecca Anthopolos, Duke University

Researchers in health and social sciences often wish to examine spatial patterns in two or more related outcomes. Examples include infant birth weight and gestational age, psychosocial and behavioral indices, and educational test scores from different topic areas. We propose a multivariate spatial mixture model for the joint analysis of correlated continuous outcomes. The responses are modeled as a finite mixture of multivariate normals, which accommodates a wide range of marginal response distributions and allows investigators to examine covariate effects across subpopulations of interest. The model has a hierarchical structure that incorporates individual- and areal-level predictors as well as spatial random effects for each mixture component. Conditional autoregressive (CAR) priors on the random effects allow the shape of the multivariate distribution to vary flexibly across geographic regions. For posterior computation, we develop an efficient Markov chain Monte Carlo (MCMC) algorithm that relies primarily on closed-form full conditionals. We use the model to explore geographic patterns in end-of-grade math and reading scores among school-age children in North Carolina.

email: brian.neelon@duke.edu

7n. ELLIPTIC SPATIAL SCAN STATISTIC ON TRENDS

Jun Luo, Information Management Services, Inc.*

Detecting clusters of heterogeneous trends of cancer rates are very useful for tracking changes of cancer status of geographic units. This paper proposes an elliptic shaped spatial scan statistic for detecting geographic clusters of trends of cancer rates. Power comparison is conducted between the elliptic shaped scan statistic and circular shaped scan statistic. The elliptic shaped scan statistic is illustrated on trends of cancer mortality rates of counties in California of US.

email: jluo.bluesky@gmail.com

7o. JOINT MODELING OF LONGITUDINAL HEALTH PREDICTORS AND CROSS-SECTIONAL HEALTH OUTCOMES VIA MEAN AND VARIANCE TRAJECTORIES

Bei Jiang, University of Michigan*
Mike Elliot, University of Michigan
Mary Sammel, University of Pennsylvania
Naisyin Wang, University of Michigan

Growth mixture models (GMMs) can be used to model the heterogeneity in the longitudinal trajectories that cannot be fully explained by measured covariates in the linear mixed effects models when assuming a common within-subject variance parameter. We can further allow subject-level variability in GMMs to differ by classifying them into different classes. As Carroll (2003) noted, "systematic dependence of variability on known factors" may be "fundamental to the proper solution of scientific problems" in certain settings, heterogeneity in different variance classes may be also important in explaining disease risks. We develop a method that simultaneously examines the association between the heterogeneities in both the mean growth profile class and the within-subject variance class and the cross-sectional binary health outcome. We consider an application to predict severe hot flashes using the hormone levels collected over time for women in menopausal transition from Penn Ovarian Aging Study.

email: beijiang6@gmail.com

7p. CONDITIONAL MAXIMUM LIKELIHOOD ESTIMATION IN TUMOR GROWTH MODELS UNDER VOLUME ENDPOINT CENSORING

Kingshuk Roy Choudhury, Duke University*
Finbarr O'Sullivan, University College Cork, Ireland

Measurements in tumor growth experiments are stopped once the tumor volume exceeds a preset threshold: a mechanism we term volume endpoint censoring. We demonstrate that this type of censoring is informative. Further, least squares parameter estimates are shown to suffer a bias in a general parametric model for tumor growth with IID measurement error, both theoretically and in simulation experiments. In linear growth models, the magnitude of bias increases with the growth rate and the standard deviation (SD) of measurement error. We propose a conditional maximum likelihood (ML) estimation procedure, which is shown both analytically and in simulation experiments to yield approximately unbiased parameter estimates. Both LS and ML estimators have similar variance characteristics. In simulation studies, these properties appear to extend to the case of moderately dependent measurement error. The methodology is illustrated by application to a tumor growth study for an ovarian cancer cell line.

email: kingshuk@duke.edu

7q. AN ANALYTICAL FRAMEWORK FOR HPV TRANSMISSION USING LONGITUDINAL DATA ON COUPLES

Xiangrong Kong, Johns Hopkins Bloomberg School of Public Health*

HPV is a common STI with 14 known oncogenic genotypes causing anogenital carcinoma. While gender-specific infections have been well studied, one remaining uncertainty in HPV epidemiology is HPV transmission within couples. Understanding transmission in couples however is complicated by the multiplicity of genital HPV genotypes and sexual partnership structures that lead to complex multi-facet correlations in data generated from HPV couple cohorts, including inter-genotype, intra-couple, and temporal correlations. We develop a mixed modeling approach using conditional probability and pairwise likelihood for analysis of longitudinal HPV couple cohort data to identify risk factors associated with HPV transmission, estimate difference in risk between male-to-female and female-to-male HPV transmission, and compare genotype-specific transmission risks within couples. The method is applied on the motivating HPV couple cohort data collected in the male circumcision trial in Rakai, Uganda to identify modifiable risk factors (including male circumcision) associated with HR-HPV transmission within couples. Knowledge from this analysis will contribute to the public health effort in preventing oncogenic HPV and related cancers in sub-Saharan Africa.

email: xikong@jhsph.edu

■ = Student Award Winner | ★ = Presenter

7r. DIMENSION REDUCTION TECHNIQUES IN APPLICATION TO LONGITUDINAL DATA ANALYSIS

Tamika Royal-Thomas, Winston-Salem State University*
Daniel McGee, Florida State University
Debajyoti Sinha, Florida State University
Clive Osmond, University of Southampton
Terrence Forrester, University of the West Indies

Longitudinal data arise when multiple observations are made on the same unit of analysis over time. This comes with the characteristics of correlated data within subjects and hence one of the measures is how do we adjust for this over time. There have been many traditional statistical methods to combat this correlation issue. Another issue that may arise is one where many variables that are correlated are collected over time and adjustments have to be made to account for this also. We examine a unique data set known as the Vulnerable Windows Data which involves three longitudinal processes collected over time. The data consists of "in utero" and early childhood data which consists of variables that are highly correlated. The application of dimension reduction techniques such as principal component analysis is utilized in creating a smaller dimension of uncorrelated data which is then utilized in a longitudinal analysis setting. This work examines how "in utero" and early childhood attributes may predict the future cardiovascular health of children as it relates to developmental origins of Adult disease. This study also examines finding an optimal linear mixed model while adjusting for both correlation within subjects over time and correlation between variables.

email: tamika.royalthomas@gmail.com

7s. A BAYESIAN SEMI-PARAMETRIC JOINT MODELING FOR LONGITUDINAL AND SURVIVAL DATA

Julius S. Ngwa, Boston University*
L. Adrienne Cupples, Boston University

Joint modeling for longitudinal and survival data has received much attention in statistical research. In previous studies these two types of data are often analyzed separately. Including the raw longitudinal measurements in the survival analysis may lead to bias if they are related to the censoring process. The analysis also becomes more complex when covariates are measured with error. We present a Bayesian semi-parametric joint model (BSJM) which links longitudinal trajectories to survival. Individual trajectories are modeled linearly and a scalar parameter is used to link the trajectories to the hazard function. We compare the BSJM approach to maximum likelihood, time dependent covariate modeling and pooled repeated observation methods. We simulate a number of scenarios and assess performance of these methods using bias, accuracy and coverage probabilities. We consider data from the Framingham Heart Study in which triglycerides (TG) measurements and Myocardial Infarction (MI) was collected over a period of 30 years (n=2306). Joint modeling indicates that individual trajectories generally increase over time (slope for log TG = 0.025 per year, 95% credible interval = (0.023, 0.028)) and are associated with risk of MI (hazard ratio = 2.97 per unit increase in log TG, 95% credible interval = (2.21, 3.86)).

email: ngwaj@bu.edu

8. MULTIVARIATE, NON-PARAMETRIC AND SEMI-PARAMETRIC MODELS

8a. ESTIMATION OF KENDALL'S TAU FOR BIVARIATE SURVIVAL DATA WITH TRUNCATION

Hong Zhu, The Ohio State University*

In many biomedical follow up studies that involve cross-sectional sampling, truncated bivariate survival data arise when one or both components of failure times is observed only if it falls within a specific truncation set. Kendall's tau is among the most popular measures of association between two random variables. For bivariate right-censored data, several nonparametric estimators have been developed by Oakes (1982), Wang and Wells (2000) and Lakhal et al. (2009). Little work has been done, however, for the nonparametric estimation of tau under both truncation and censoring. Additional selection bias due to truncation effect must be adjusted for in comparing a pair of observed failure times. This paper takes the approach of inverse probability weighting to account for both comparability and orderability of pairs. Nonparametric estimators of tau for measuring the association are proposed for bivariate left-truncated data, and bivariate survival data with interval sampling where one component is subject to double truncation and the other is subject to possibly dependent right censoring. The estimators are shown to be consistent and asymptotically normally distributed. The methods developed are applicable to many patterns of truncation for bivariate survival data, including left-truncation, right truncation and double truncation for one or both components, where no other nonparametric estimator of tau is currently available. Simulation studies demonstrate that the proposed estimators perform well with moderate sample size. A real data example is provided for illustration of the methods and theory.

email: hzhu@cph.osu.edu

8b. PROBABILISTIC INDEX MIXED MODELS FOR CLUSTERED DATA

Fanghong Zhang, Ghent University, Belgium*
Stijn Vansteelandt, Ghent University, Belgium and London School of Hygiene and Tropical Medicine, U.K.
Jan De Neve, Ghent University, Belgium
Olivier Thas, Ghent University, Belgium

The use of linear mixed models for continuous clustered data can be problematic when the data distribution is skewed in a way that is not easily accommodated by data transformation, or when the response is measured on an ordinal scale. The probabilistic index models (PIM) of Thas et al. (2012) form a flexible class of semi-parametric models for analyzing non-normal responses, but they assume mutual independence of all observations. In view of this, we propose an extension of the PIMs to clustered data. Given two random vectors (Y_{ik},X_{ik}) and (Y_{jl},X_{jl}), where Y_{ik} and X_{ik} denote the kth outcome and covariate measurement for subject i, we consider models of the form: P(Y_{ik} <= Y_{jl}|X_{ik},X_{jl},b_i,b_j) =\Phi{beta*(X_{ik}-X_{jl})+(b_i-b_j)} Where b_i and b_j are i.i.d random effects with constant variance, beta is a finite-dimensional parameter and \Phi(.) is the standard normal cumulative distribution function. Consistent and asymptotically normal estimators of the covariate effect beta and of the random effects variance are established through semi-parametric estimation methods. The proposed framework extends non-parametric rank tests to deal with clustered data in settings that require covariate adjustment. In a simulation study, the finite-sample behavior of our estimators is evaluated and a data analysis example is provided.

email: fanghong.zhang@ugent.be

8c. A MULTI-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 there are multiple p-values associated with each test. For example, we see such p-values in the study of cell-cycle regulated genes of the yeast Saccharomyces cerevisiae. In a study by Spellman et al (Mol. Biol. Cell., 1998), yeast cells were synchronized using three independent methods and thus gave rise to up to three distinct p-values to test periodicity for each gene. In this work we explore an approach that accounts for the multi-dimensional spatial structure of these vectors of 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 hypercube. We explore various ordering schemes to rank the p-vectors, and use an empirical null approach to control the false discovery rate. We use our approach to analyze the Saccharomyces cerevisiae data and compare our results to previous studies.

email: daisylahaina@gmail.com

8d. MODEL SELECTION AND ESTIMATION IN GENERALIZED ADDITIVE MIXED MODELS

Dong Wang, North Carolina State University*
Daowen Zhang, North Carolina State University

We propose a method of model selection and estimation in generalized additive mixed models (GAMMs) where subject-specific random effects are introduced to accommodate the correlation among measurements. The linear mixed model representation of the smoothing spline estimators of the nonparametric functions is used, where the importance of these functions is controlled by treating the inverse of the smoothing parameters as extra variance components. By maximizing the penalized likelihood with the adaptive LASSO, we could estimate the variance components from the nonparametric functions and therefore select the important ones. In addition, a unified EM algorithm is provided to obtain both the maximum likelihood and maximum penalized likelihood estimates of the variance components. In order to overcome the computational problems due to the large sample sizes in longitudinal data, we use the eigenvalue-eigenvector decomposition method to approximate the working random effects. In this case, the dimensions of matrices in the algorithm are greatly reduced while keeping most data information.

email: dwang4@ncsu.edu

8e. CLINICAL VARIABLES ASSOCIATED WITH MELANOMA BRAIN METASTASIS: A META ANALYSIS

Meng Qian, New York University School of Medicine*
Michelle Ma, New York University School of Medicine
Ronald O. Perelman, New York University School of Medicine
Iman Osman, New York University School of Medicine
Yongzhao Shao, New York University School of Medicine

Introduction: Brain metastatic (B-met) melanoma patients have a very high mortality rate with a short median survival time of less than 6 months. It is critically important to identify among early-stage primary melanoma patients who have high risk to metastasize to the brain. The melanoma research team (IMCG) at NYU School of Medicine has acquired a unique resource with three prospectively collected melanoma databases over several decades. Methods and Results: After Literature search, only nine articles provided both case (B-met) and control (none-B-met) information and were identified as suitable for meta-analysis in assessing risk factors that are of potentially predictive value for B-Met development. In the meta analysis, the summary odds ratios with corresponding 95% confidence interval (95% CIs)

were estimated by pooling the study-specific estimates using the random effects model. We found that several clinical factors are statistically significant risk factors for melanoma B-met. The effects of clinical factors and demographic variables on B-Met development in two melanoma databases were further assessed by multiple logistic regression analysis.

email: mq289@nyu.edu

8f. RANK BASED ESTIMATION FOR GENERALIZED LINEAR MODELS

Guy-Vanie M. Miakonkana, Auburn University*
Asheber Abebe, Auburn University

In this paper we consider the estimation of parameters of a generalized linear regression model. An estimator defined iteratively, starting from an initial obtained by minimizing the Wilcoxon dispersion function for independent errors, is considered. The consistency and the asymptotic normality of the initial estimator as well as the asymptotic normality of the updated estimator are proved under minimal assumptions. Like in linear model, the procedure results in estimators that are robust in the response space. We present results of a simulation study as well as real world data example to illustrate the robustness and efficiency of the estimator.

email: gmm0006@auburn.edu

8g. EXPLORING MULTIVARIATE ASSOCIATIONS: A GRAPH THEORETIC APPROACH REVISITED

Srikesh G. Arunajadai, Columbia University*

We revisit the graph theoretic approach to multivariate associations using inter-point distance based graphs. The graphs used are k-Minimal Spanning Trees (k-MST) and k-Nearest Neighbor graphs (k-NN). We provide ways to choose k to satisfy the distributional assumptions on the test statistic. The method does not assume any underlying distribution for the data. It is often of interest to understand associations between vector-valued variables. For example, in the behavioral sciences, to study the association between batteries of cognitive tests differing in type and number of tests administered at different ages. In functional data analysis, it may be of interest to assess association between functional covariates. In high dimensional data sets, statistics to test the null hypothesis of independence between possibly vector valued random variables of arbitrary dimensions against a broad set of alternatives can be beneficial in screening relevant covariates for further analysis. Recently, tests and measures of multivariate associations have been proposed using copulas and distance correlations. Simulations are performed to compare the commonly used multivariate tests of association with the graph-based methods. We provide conditions under which the tests help in identifying associations.

email: sa2658@columbia.edu

8h. COMPARISON OF RANK BASED TESTS IN COMBINED DESIGNS

Yvonne M. Zubovic, Indiana University Purdue University Fort Wayne*

A simulation study was conducted to investigate the performance of variance statistics for testing differences of effects when k treatments are administered. One constraint under consideration is that data are to be utilized from two experimental designs: a one-way layout and a completely randomized block design. A second constraint is that Normality assumptions are not met. In this study several nonparametric methods based on ranks were considered. Using simulation we compare the performance of tests based on aligned ranks, a rank transformation, and a linear combination of nonparametric tests when sample sizes are not balanced and variances differ under various underlying distributional shapes. The properties of the proposed hypothesis tests are shared.

email: zubovic@ipfw.edu

8i. FACTOR ANALYSIS FOR BINARY DATA USED TO MEASURE PATIENT'S EXPECTATIONS AND EXPERIENCES WITH HEALTH SERVICES

Rebeca Aguirre-Hernandez, Ciudad Universitaria, Mexico*
Alicia Hamui-Sutton, Ciudad Universitaria, Mexico
Ruth García-Fuentes, Ciudad Universitaria, Mexico
Anselmo Calderon-Estrada, Ciudad Universitaria, Mexico

A questionnaire was designed to assess the service provided by healthcare institutions in Mexico and to measure patients' satisfaction. Many surveys seek to determine if healthcare institutions met their goals, our questionnaire investigated the patient´s point of view. 2176 patients attended in the ambulatory, specialized and emergency departments of 19 different hospitals in Mexico were interviewed at the end of their visit. The information was collected on November 2010. Patients were asked if they had any economic, work, social or emotional concerns before receiving attention. They were also asked how the visit to the doctor did and the treatment prescribed affected their economic, work, social and emotional life. We inquired about patient´s socio-demographic characteristics, health fitness, and moral values and investigated about sources of satisfaction and dissatisfaction with regard to the hospital premises, the physicians, and other staff. We present the results of a factor analysis for binary data used to create indexes that reflect patient's expectations and experiences with regard to their economic, work, social and emotional life.

email: rebecaaguirrehdez@yahoo.com.mx

8j. AUTOASSOCIATIVE NEURAL NETWORK APPROACH FOR NONLINEAR PRINCIPAL COMPONENT ANALYSIS

Siddik Keskin, Yuzuncu Yil University and University of Toronto*
W.Y. Wendy Lou, University of Toronto

Nonlinear principal component analysis (NPCA), a nonlinear generalization of standard principal component analysis, can be implemented using a neural network, such as the auto associative neural network (AANN), for feature extraction and dimensionality reduction. In this study, we present an AANN approach to NPCA, followed by an application to real data from a survey study of university students in Turkey to examine the relationships between depression and various risk factors. The effectiveness and applicability of the approach will be discussed, and some recommendations will be given.

email: skeskin973@gmail.com

9. MODELING, PREDICTION, DIAGNOSTIC TESTING, VARIABLE SELECTION AND CONSULTING

9a. JOINT MODELING OF CENSORED MULTIVARIATE LONGITUDINAL AND EVENT TIME DATA

Francis Pike, University of Pittsburgh*
Lisa Weissfeld, University of Pittsburgh

In the biomedical sciences it is often of interest to evaluate the prognostic value of one or more potential biomarkers and their joint association with a survival endpoint. In the setting considered herein we were interested in the joint association of two heavily censored longitudinal markers of inflammation, IL6 (Interleukin-6) and IL10 (Interleukin-10), with survival. In this scenario, the analytic method has to simultaneously account for heavy censoring in one or more of the longitudinal measures and for the inherent dependence between such measures. To account for the longitudinal censoring and dependence between biomarkers we developed a multivariate Joint Tobit Model whereby we linked the longitudinal evolutions of two, possibly, dependent censored markers of inflammation to survival by assuming that the hazard was a function of both of the longitudinal trajectories with shared random effects. We validated this approach via simulation and then applied the method to the GENIMS study (Genetic Markers of Inflammation Study) data where both of the aforementioned analytical challenges needed to be addressed.

email: frp3@pitt.edu

9b. CROSS-SECTIONAL HIV-1 INCIDENCE ESTIMATION UTILIZING VIRAL GENETIC DIVERSITY

Natalie M. Exner, Harvard School of Public Health*
Vladimir A. Novitsky, Harvard School of Public Health
Marcello Pagano, Harvard School of Public Health

The ability to accurately estimate human immunodeficiency type 1 (HIV-1) incidence is critical for understanding transmission dynamics, but existing methods for incidence estimation have extensive limitations. Bioassays, such as those that exploit the maturation of the host immune system, provide a cross-sectional test of recency of infection, making them more affordable and less prone to bias than longitudinal follow-up of a cohort. Nonetheless, these immunoassays have been shown to have subtype variability, misclassify late-stage infections, and misclassify those on antiretroviral treatment. An emerging alternative to immunoassays are assays which measure viral diversity within a host. Using data from a longitudinal cohort of subtype C-infected persons from Botswana, we have developed a novel method of incidence estimation which incorporates measures of within-host viral diversity. We will describe the different components of our testing algorithm and its preliminary performance characteristics.

email: nexner@hsph.harvard.edu

9c. JACKKNIFE EMPIRICAL LIKELIHOOD FOR ROC CURVES WITH MISSING DATA

Hanfang Yang, Georgia State University*

In this paper, we apply jackknife empirical likelihood (JEL) method to construct confidence intervals for the receiver operating characteristic (ROC) curve with missing data. After using hot deck imputation, we generate pseudo-jackknife sample to develop jackknife empirical likelihood. Comparing to traditional empirical likelihood method, JEL has a great advantage in saving computational cost. Under mild conditions, the jackknife empirical likelihood ratio converges to a scaled chi-square distribution. Furthermore, simulation studies in terms of coverage probability and average length of confidence intervals demonstrate this proposed method have the good performance in small sample sizes.

email: hyang13@student.gsu.edu

■ = Student Award Winner | ★ = Presenter

9d. PROFILE LIKELIHOOD BASED CONFIDENCE INTERVAL OF THE INTRACLASS CORRELATION FOR BINARY OUTCOMES, WITH APPLICATIONS TO TOXICOLOGICAL DATA

Krishna K. Saha, Central Connecticut State University*

The intraclass correlation in binary outcome data is an important and versatile measure in many biological investigations. Properties of the different estimators of the intraclass correlation based on different approaches have been studied extensively, mainly, in light of bias and efficiency, but little attention has been paid in the extension of these results to the problem of the confidence intervals of it. In this article, we extend the results of the four point estimators by constructing asymptotic confidence intervals obtaining closed-form asymptotic and sandwich variance expressions of those four point estimators. It appears from simulation results that the asymptotic confidence intervals based on these four estimators have serious under-coverage. To remedy this, we introduce the profile likelihood approach based on the beta-binomial model and the hybrid profile variance approach based on the quadratic estimating equation for constructing the confidence intervals of the intraclass correlation for binary outcome data. As assessed by simulations, the proposed confidence interval approaches show significant improvement in the coverage probabilities. Moreover, the profile likelihood approach performs quite well by providing coverage levels close to nominal over a wide range of parameter combinations. Application to toxicological data is provided to illustrate the methods.

email: sahakrk@ccsu.edu

9e. INTRODUCTORY STATISTICS FOR MEDICAL STUDENTS---IN 6 LECTURES

Jacob A. Wegelin, Virginia Commonwealth University*

Incoming medical students at Virginia Commonwealth University (n=200) are required to complete a 34-hour, team-taught course in Population Medicine, of which a biostatistician is responsible for 6 hours. No prior statistical training is assumed. Non-statisticians in charge of the course presented me with a list of topics to cover that would have required a semester. They requested that I freeze all content two months in advance by publishing powerpoints, hard copies of which the students would leaf through during lecture instead of taking notes. I rejected this approach, and focused instead on old-fashioned lectures using a whiteboard. Rather than attempt to cover a semester's material superficially, I re-designed the course to focus on a few essential concepts. I hoped that the students would grasp (1) the notions of a dataset, variable types, and empirical distributions, (2) the duality of the empirical and theoretical worlds, rooted in Plato's philosophy, and (3) the frequentist notions of p value and confidence interval. Instead of powerpoints, I wrote an 87-page text. Lectures were not shackled to the text. I will sketch the process, report the outcome, and suggest future work.

email: JacobWegelin@FastMail.FM

9f. COMPARISON OF TESTS IN A REGION AROUND THE OPTIMAL THRESHOLD

Donna K. McClish, Virginia Commonwealth University*

There are 3 primary ways to compare two medical tests via their corresponding ROC curves. These include 1) determining whether ROC curves are exactly the same, a global test which requires that true positive rates (TPRs) be the same at all false positive rates (FPRs); 2) comparing ROC curves at a particular FPR, which implies that there is a single, specific FPR of interest; or 3) comparing the area or partial area under the ROC curve. The full area implies interest in all FPRs, while the partial area requires choice of a subset of FPRs. We suggest comparing tests at the optimal point, defined as the point where the Youden Index is maximized. But rather than compare curves at this single point, we suggest that a better choice would be to compare the portion of the ROC curves centered about the optimal point. This portion of the ROC curve could be defined as the confidence interval (CI) around the optimal point. For binormal data, Schisterman et al (2007) and Skaltsa et al (2010) provide formulae for the CIs of the optimal threshold and Youden Index. CIs for the partial area must take into account that the limits of integration are random rather than fixed.

email: mcclish@vcu.edu

9g. INTERQUANTILE SHRINKAGE IN REGRESSION MODELS

Liewen Jiang, North Carolina State University*
Howard Bondell, North Carolina State University
Judy Wang, North Carolina State University

Conventional research on quantile regression often focuses on fitting the regression model at different quantiles separately. However, in situations where the quantile coefficients share some common feature, joint modeling of multiple quantiles to accommodate the commonality often leads to more efficient estimation. One example of common feature is that for some predictors, the quantile coefficients are constant in some regions of the quantile level, but vary in other regions. To automatically perform estimation and detection of the interquantile commonality, we develop two penalization methods. When the quantile slope coefficients indeed do not change across quantile levels, the proposed methods will shrink the slopes towards constants and thus improve the estimation efficiency. We established the oracle properties of the two proposed penalization methods. Through numerical investigations, we demonstrate that the proposed methods lead to estimations with competitive or higher efficiency than the standard quantile regression estimation in finite samples.

email: ljiang2@ncsu.edu

9h. A JOINT MODEL FOR QUALITY OF LIFE AND SURVIVAL IN PALLIATIVE CARE STUDIES

Zhigang Li, Dartmouth Medical School*
Tor Tosteson, Dartmouth Medical School
Marie Bakitas, Dartmouth Medical School

Palliative care is a specialized area of healthcare that focuses on helping relieve and prevent the suffering of patients, especially those patients facing life-threatening illness. These services are available in most (80%) large U. S. hospitals. Survival data and longitudinal measurements of quality of life (QoL) are usually collected in palliative care studies, where palliative care is compared with usual care. Models and inferential methods have been proposed for jointly analyzing the longitudinal outcome processes and event times, primarily from the perspective of corrections for missing data and the dependence of survival on surrogate markers of disease progression. We propose a terminal decline model focusing on the decline of QoL during the last months of life. Two sub models are used for the longitudinal outcomes (mixed model) and survival times (piecewise exponential). In the mixed model, the time scale is counting backward from death so that we are able to directly compare QoL experienced at fixed times from death. Non-informative censoring is assumed in the model since death times can be found out at the termination of the study. Explicit formulae of the quality-adjusted life years can be derived and compared between the treatment groups.

email: zhigang.li@dartmouth.edu

9i. THE HOSMER-LEMESHOW GOODNESS OF FIT TEST FOR MULTIPLY IMPUTED DATA

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

The Hosmer-Lemeshow (H-L) test is widely used for evaluating goodness of fit in logistic regression models. The H-L test first creates groups based on deciles of estimated probabilities and then compares observed and expected event rates within these groups. Multiple imputation (MI) is growing in popularity as a method for handling missing data, and how to apply the H-L test after MI is not straightforward. In this paper we discuss complexities involved in applying the H-L test to multiply imputed data, related to which variables have missingness. When covariates have been imputed, predicted probabilities vary across imputed data sets, and thus the boundaries of the predicted probability groupings vary as well. When the outcome has been imputed, both predicted probabilities and observed event rates vary across MI data sets. We then propose several different methods for using the H-L test with multiply imputed data, and compare the methods through simulation.

email: randridge@cph.osu.edu

9j. REVERSE KAPLAN-MEIER METHOD FOR ANALYZING BIOMARKERS WITH LIMIT OF DETECTION

Tulay Koru-Sengul, University of Miami*

Biomarkers are laboratory measures of biological processes. Because they are viewed to be objective and quantifiable by providing valuable information in both assessing exposure and disease status, they are increasingly used in biomedical and public health sciences, drug development programs, testing for targeted therapeutics and personalized medicine. This increased use creates a venue for the development of methods to address important, unexplored analytic issues such as properly handling biomarker measurements below the limit of detection (LOD). Researchers working with biomarker data inevitably have to deal with data containing non-detects, and how to combine non-detects with values above the LOD for data analysis. The common statistical analysis approach has been to use ad-hoc single imputation techniques then conduct the analysis under the assumption that the imputed values are the actual observed values. The Kaplan-Meier method is a better alternative to these ad-hoc single imputation techniques. We will facilitate broader use of the reverse Kaplan-Meier estimator by describing its properties, illustrating its use with population-based data for secondhand smoke exposure research and showing how it can be calculated using standard software. Results from Monte-Carlo simulation studies will be used to demonstrate the method in different scenarios. Funding provided by FAMRI, NIH, NIOSH.

email: tsengul@med.miami.edu

email: tsengul@med.miami.edu

9k. EMPIRICAL LIKELIHOOD BASED TESTS FOR STOCHASTIC ORDERING IN RIGHT-CENSORED SETTING

Hsin-wen Chang, Columbia University*
Ian W. McKeague, Columbia University

This talk introduces a new empirical likelihood approach to testing for the presence of stochastic ordering between two univariate distributions when the data are right-censored. The proposed test statistic is the supremum of a localized, weighted, empirical likelihood statistic. The asymptotic null distribution of the test statistic is derived in terms of Brownian motion, and simulation is used to obtain critical values. Applications of the proposed test to randomized clinical trials with time-to-event endpoints are discussed.

email: hc2496@columbia.edu

9l. EFFICIENT ESTIMATION USING CONDITIONAL EMPIRICAL LIKELIHOOD WITH MISSING OUTCOMES

Peisong Han, University of Michigan*
Lu Wang, University of Michigan
Peter X.K. Song, University of Michigan

In many biomedical studies, the outcome is measured only on part of study subjects, with some surrogate correlated with the outcome available for all subjects. We consider regression model of the mean of the outcome on covariates indexed by finite number of regression parameters, and assume the missing outcomes are missing at random. Estimation based on the efficient score presents a great challenge, as it requires to model the second moment of an augmented inverse probability weighted (AIPW) residual, which involves several random processes. We propose a conditional empirical likelihood (CEL) approach to overcome this difficulty. The CEL estimation relies solely on the AIPW residual, with no need to explicitly model the second moment. We show the CEL estimator enjoys the double robustness property, i.e., it is consistent if either the missing probability or the conditional mean of the outcome given the surrogate and covariates is correctly modeled. When both models are correct, the CEL estimator achieves the semiparametric efficiency bound. We also establish the asymptotic normality when either one model is correct. Simulation studies are conducted to evaluate the finite sample performance.

email: peisong@umich.edu

9m. A PERTURBATION METHOD FOR PREDICTION ACCURACY WITH REGULARIZED REGRESSION

Jessica Minnier, Harvard School of Public Health
Tianxi Cai, Harvard School of Public Health*

Analysis of massive 'omics' data often seeks to identify a subset of genes or proteins that are predictive of disease outcomes. Traditional statistical methods for variable selection often fail in the presence of high-dimensional features. Classification algorithms based on genetic and biological markers have been developed for prediction of clinical outcomes. Robust regularization methods can achieve an optimal trade-off between the complexity of the model by simultaneously performing variable selection and estimation, leading to more accurate prediction models. However, in finite samples, it remains difficult to evaluate the predictive performance of such models. Estimates of accuracy measures such as absolute prediction error, ROC curves, and AUC statistics may be imprecise, especially in the small study setting when cross-validation procedures are used, and therefore model comparison is challenging. We propose perturbation resampling based procedures to approximate the distribution of such prediction accuracy measures in the presence of regularized estimation. This method provides a simple way to estimate confidence regions for the true prediction accuracy of a model. Through finite sample simulations, we verify the ability of our method to accurately evaluate the prediction model and compare it to other standard methods. We also illustrate our proposals with a study relating HIV drug resistance to genetic mutations.

email: jminnier@hsph.harvard.edu

10. STATISTICAL GENOMICS IN SEQUENCING ERA, FROM DATA ANALYSIS TO PERSONAL MEDICINE

QUANTITATIVE TRAIT ANALYSIS UNDER TRAIT-DEPENDENT SAMPLING, WITH APPLICATIONS TO THE NHLBI EXOME SEQUENCING PROJECT

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

In the NHLBI Exome Sequencing Project, 267 subjects with BMI values >40 and 178 subjects with BMI values <25 (and without diabetes) were selected for sequencing out of a total of 11,468 subjects from the Women's Health Initiative. Similar designs were used for LDL and blood pressures, although the sampling was based on age- and gender-adjusted residuals rather than raw measurements. Because the sampling depends on the trait values, standard analysis methods, such as least-squares estimation, are not appropriate. We have developed valid and efficient statistical methods for quantitative trait analysis under complex trait-dependent sampling. Our methods can be used to perform quantitative trait analysis not only for the trait that is used to select the subjects for sequencing but also for any other quantitative traits that are measured. For any particular quantitative trait, the association results from all available samples are combined through meta-analysis. The efficiency gain of such meta-analysis over the analysis of each trait based on its own sample alone is enormous and can lead to many more discoveries. The advantages of the new methods over the existing ones are demonstrated through simulation studies and analysis of data from the NHLBI Exome Sequencing Project. The relevant software is freely available.

email: lin@bios.unc.edu

A SURVIVAL-SUPERVISED LATENT DIRICHLET ALLOCATION MODEL FOR GENOMIC BASED STUDIES OF DISEASE

John A. Dawson, University of Wisconsin - Madison
Christina Kendziorski, University of Wisconsin - Madison*

Statistical methods for variable selection, prediction, and/or classification have proven extremely useful in moving personalized genomic medicine forward, in particular leading to a number of genomic based assays now in clinical use. Although invaluable in individual cases, the information provided by these assays is limited. In particular, a patient is often classified into one of very few groups (e.g. recur or not), limiting the potential for truly personalized treatment. Furthermore, although these assays provide information on whether or not to treat (e.g. if recurrence is predicted), they provide no information on how to treat. This talk presents a survival-supervised latent Dirichlet allocation (LDA) model developed to address these limitations. In particular, we extend the traditional LDA framework to the genome so that distinct topics (collections of genomic aberrations, clinical variables, and treatments) can be derived and the contribution of each topic to a document (the description of an individual's genome and phenome, treatment regime, and disease course) can be determined. The framework facilitates data integration across multiple platforms and scales to enable powerful patient-specific inference. Advantages of the approach are illustrated using data from the Cancer Genome Atlas (TCGA) project.

email: kendzior@biostat.wisc.edu

DETECTION OF RNA AND DNA SEQUENCE DIFFERENCES IN THE HUMAN TRANSCRIPTOME

Mingyao Li, University of Pennsylvania*

The transmission of information from DNA to RNA is a critical process. In our recent paper, we compared RNA sequences from human B cells of 27 individuals to the corresponding DNA sequences from the same individuals and uncovered more than 10,000 exonic sites where the RNA sequences do not match that of the corresponding DNA. All 12 possible categories of discordance were observed. These differences were nonrandom as many sites were found in multiple individuals and in different cell types. These widespread RNA-DNA differences (RDDs) in the human transcriptome provide an unexplored aspect of genomic variation. While encouraging, these results were based on the analysis of single individuals with hard coded threshold. In this talk, I will present a likelihood-based approach to detect RDDs by combining information across multiple individuals, incorporating sequencing errors as well as information on SNPs. I will show that such multi-sample based calling algorithm improves both calling sensitivity and specificity. I will also describe a statistical framework that treats RDDs as a type of genomic variants in downstream gene mapping analysis. Such approach is complementary to the typical SNP association at the DNA level and may uncover novel variations of biological and clinical relevance.

email: mingyao@mail.med.upenn.edu

ESTIMATION OF SEQUENCING ERROR RATES IN SHORT READS

Xin Victoria Wang, Dana-Farber Cancer Institute; Harvard School of Public Health*
Natalie Blades, Brigham Young University
Jie Ding, Dana-Farber Cancer Institute; Harvard School of Public Health
Razvan Sultana, Dana-Farber Cancer Institute; Boston University
Giovanni Parmigiani, Dana-Farber Cancer Institute; Harvard School of Public Health

Short-read data from next-generation sequencing technologies are now being generated across a range of research projects. The fidelity of this data can be affected by several factors and it is important to have simple and reliable approaches for monitoring it at the level of the individual experiment. We developed a fast, scalable and accurate approach to estimating error rates in short reads, which has the added advantage of not requiring a reference genome. We build on the fundamental observation that there is a linear relationship between the copy number for a given read and the number of reads that differ from theread of interest by one or two bases. The slope of this relationship can be transformed to give an estimate of the error rate, both by read and by position. We present simulation studies as well as analyses of real data sets illustrating the precision and accuracy of this method, and we show that it is more accurate than alternatives that count the difference between the sample of interest and the reference genome. We show how this methodology led to the detection of mutations in the genome of the PhiX strain used for normalization of Illumina data.

email: vwang@jimmy.harvard.edu

11. VARIABLE SELECTION FOR COMPLEX MODELS

A ROAD TO CLASSIFICATION IN HIGH DIMENSIONAL SPACE

Jianqing Fan, Princeton University*
Yang Feng, Columbia University
Xin Tong, Princeton University

For high-dimensional classification, researchers proposed independence rules to circumvent the diverse spectra, and sparse independence rule to mitigate the issue of noise accumulation. However, in biological applications, there are often a group of correlated genes responsible for clinical outcomes, and the use of the covariance information can significantly reduce misclassification rates. The extent of such error rate reductions is unveiled by comparing the misclassification rates of the Fisher discriminant rule and the independence rule. To materialize the gain based on finite samples, a Regularized Optimal Affine

Discriminant (ROAD) is proposed based on a regularized Fisher discriminant. ROAD selects an increasing number of features as the penalization relaxes. Further benefits can be achieved when a screening method is employed to select a subset of variables. A constrained coordinate-wise descent (CCD) algorithm is also developed to solve the optimization problem related to ROAD. Oracle type of sampling properties are established. Simulation studies and real data analysis support our theoretical results and demonstrate the advantages of the new classification procedure under a variety of correlation structures.

email: jqfan@princeton.edu

BAYESIAN NONPARAMETRIC VARIABLE SELECTION

David Dunson, Duke University*

There is an increasingly vast literature on variable selection in parametric models, such as linear regression and GLMs, but little attention has been focused on nonparametric Bayes variable selection. In this talk, I start by considering general mean regression problems involving p candidate predictors, so that interest focuses on estimating a p-dimensional unknown regression surface. We propose classes of Bayesian nonparametric variable selection procedures relying on scaled Gaussian processes and tensor product specifications. Sufficient conditions are provided under which the posterior achieves an adaptive minimax rate of convergence as if the true variables and smoothness of the regression surface were known. We then propose a new class of Bayesian models for conditional densities (density regression), which lead to straightforward posterior computation allowing variable selection and inherit the adaptive optimal rate results of the mean regression case. Small sample properties are evaluated through simulations and the methods are applied to epidemiology data. Joint work with Debdeep Pati and Anirban Bhattacharya.

email: dunson@stat.duke.edu

RISK PREDICTION FROM GENOME-WIDE DATA

Ning Sun, Yale School of Public Health
Hongyu Zhao, Yale School of Public Health*

Thousands of genetic variants have been found to be associated with complex traits in humans in recent years through genome wide association studies. It is likely many more regions/markers will be implicated through next generation sequencing in the near future. However, these discoveries have not been translated into informative prediction models for these traits. In this presentation, we will discuss the limitations of the current methods and propose a mixed effects model-based modeling approach for prediction that involves the selection of markers from millions of potential candidates and the assessment of overall genetic similarities among study subjects.

email: hongyu.zhao@yale.edu

COMPLETE LEAST SQUARES FOR SCREENING AND VARIABLE SELECTION

Leonard A. Stefanski, North Carolina State University*
Eric Reyes, North Carolina State University
Dennis Boos, North Carolina State University

We construct a new linear regression model object function for the purpose of screening a large number of predictors followed by variable selection. The objection function is obtained as the sum of all of objectives functions for all models of all sizes, hence the name "Complete Least Squares" (CLS). We show that ordering (by absolute value) the estimated regression coefficients derived from the complete least squares objective function produces screening sequences with favorable properties, i.e., the true nonzero coefficients tend to stack up early in the sequence. Properties of the CLS estimator and the selection sequence are studied analytically and via a Monte Carlo study.

email: stefansk@stat.ncsu.edu

12. OPTIMAL AND PERSONALIZED TREATMENT OF HIV

METHODS FOR EVALUATING THE EFFECTS OF DELAYED ARV REGIMEN CHANGE

Brent A. Johnson, Emory University*

The current goal of initial antiretroviral (ARV) therapy is suppression of plasma HIV-1 RNA levels to below the limits of currently available assays. A substantial proportion HIV-infected patients who initiate antiretroviral therapy in clinical practice or antiretroviral clinical trials either fail to suppress HIV RNA or have HIV RNA levels rebound on therapy. The optimal time to switch antiretroviral therapy to ensure sustain virologic suppression and prolonged clinical stability in patients who have rebound in their HIV RNA, yet are stable clinically, is unknown. Despite repeated attempts, randomized clinical trials to compare early versus delayed switching have been difficult to design and enroll. In many clinical trials, patients randomized to initial antiretroviral treatment combinations, who fail to suppress HIV RNA or have a rebound of HIV RNA on therapy are allowed to switch the ARV regimen to which they were randomized in favor of another regimen based on clinician and patient decisions. We propose a statistical frameworks to quantify how the timing of ARV regimen change after confirmed virologic failure affects clinical outcome. The methods are motivated by and applied to data from the AIDS Clinical Trials Group (ACTG) Study A5095.

email: bajohn3@emory.edu

PERSONALIZED MEDICINE FOR HIV PATIENTS INITIATING THERAPY

Brian Claggett, Harvard School of Public Health*
Yun Chen, Harvard School of Public Health
Michael Hughes, Harvard School of Public Health
Heather Ribaudo, Harvard School of Public Health
Camlin Tierney, Harvard School of Public Health
Katie Mollan, Harvard School of Public Health

With the increased availability of highly active antiretroviral therapy (HAART) in many regions, the prognosis for HIV patients has improved substantially and it is increasingly apparent that regimens must be both effective and tolerable for patients for successful long-term treatment. To this end, we utilize the data from three recent AIDS Clinical Trials Group (ACTG) studies to develop and validate a risk prediction model for regimen failure among HIV patients beginning treatment. In addition to investigating factors associated with regimen failure, we investigate whether the predicted risk of regimen failure can be utilized in the context of personalized treatment decisions by applying the proposed model to the patients in another recent ACTG study and assessing the treatment effects among patients with differing levels of predicted risk.

email: bclagget@hsph.harvard.edu

ESTIMATION OF CONSTANT AND TIME-VARYING DYNAMIC PARAMETERS OF HIV INFECTION IN A NONLINEAR DIFFERENTIAL EQUATION MODEL

Hulin Wu, University of Rochester*
Hongyu Miao, University of Rochester
Hua Liang, University of Rochester

Modeling viral dynamics in HIV/AIDS studies has resulted in deep understanding of pathogenesis of HIV infection from which novel antiviral treatment guidance and strategies have been derived. Viral dynamics models based on nonlinear differential equations have been proposed and well developed over the past few decades. However, it is quite challenging to use experimental or clinical data to estimate the unknown parameters (both constant and time-varying parameters) in complex nonlinear differential equation models. It is desirable to determine all the parameter estimates from data. In this study, we intend to combine the newly developed approaches, a multi-stage smoothing-based (MSSB) method and the spline-enhanced nonlinear least squares (SNLS) approach, to estimate all HIV viral dynamic parameters in a nonlinear differential equation model. In particular, to the best of our knowledge, this is the first attempt to propose a comparatively thorough procedure, accounting for both efficiency and accuracy, to rigorously estimate all key kinetic parameters in a nonlinear differential equation model of HIV dynamics from clinical data. These parameters include the proliferation rate and death rate of uninfected HIV-targeted cells, the average number of virions produced by an infected cell, and the infection rate which is related to the antiviral treatment effect and is time-varying. To validate the estimation methods, we verified the identifiability of the HIV viral dynamic model and performed simulation studies. We applied the proposed techniques to estimate the key HIV viral dynamic parameters for two individual AIDS patients treated with antiretroviral therapies. We demonstrate that HIV viral dynamics can be well characterized and quantified for individual patients. As a result, personalized treatment decision based on viral dynamic models is possible.

email: hwu@bst.rochester.edu

13. STATISTICAL METHODS FOR HOSPITAL COMPARISONS

HOSPITAL COMPARISONS, ISSUES AND APPROACHES

Arlene Ash, University of Massachusetts Medical School
Stephen E. Fienberg, Carnegie Mellon University
Thomas A. Louis, Johns Hopkins Bloomberg School of Public Health*
Sharon-Lise T. Normand, Harvard Medical School and Harvard School of Public Health
Therese Stukel, University of Toronto
Jessica Utts, University of California, Irvine

The Centers for Medicare and Medicaid Services (CMS) annually compares hospitals with respect to mortality, readmissions, and other outcomes. Performance metrics are based on the best possible information for each hospital as to how well it performs with its patients in comparison to outcomes that would be expected, if the same patients were to receive care that matched the national norm. The CMS requested the Committee of Presidents of Statistical Societies to prepare a white paper on the appropriateness of the statistical methods used for these measures. The resulting white paper addresses statistical issues associated with the hierarchical generalized linear modeling currently used in producing the CMS metrics. Using the white paper as background, we review the current CMS approaches to evaluating mortality and readmissions, outline the assumptions underlying these methodologies, and evaluate the extent to which they respect the data structures and address stated goals. Handling the low information context is especially important and we consider the consequent need for smoothing/stabilization, specifically via shrinkage estimation. We identify potential improvements to the current CMS methodology, and recommend methodological and empirical studies to inform which, if any, changes are required.

email: tlouis@jhsph.edu

ADVANTAGES OF UNIFIED MODEL-BASED APPROACHES TO PROVIDER PROFILING

Frank E. Harrell, Vanderbilt University School of Medicine*

I will discuss general advantages of model-based statistical measures of outcome differences as opposed to ad-hoc measures that were created to provide crude population-based summaries. The latter includes observed-to-expected ratios, which hospitals sometimes use to hide the drivers of time trends in outcome quality. The major point advocated will be that the ultimate goal is prediction of an individual hospital or provider's future outcomes for specific patient types, and that statistical methods that optimize predictive accuracy should be chosen over methods that are known to have higher mean squared error of forecasts. Full conditioning on covariates rather than estimating population averages, and the use of random effects are important components of optimal solutions to the provider profiling problem.

email: f.harrell@vanderbilt.edu

CHALLENGES IMPLEMENTING A HIERARCHICAL MODEL FOR USE IN PUBLIC REPORTING: THE CASE OF "HOSPITAL COMPARE"

Jeffrey H. Silber, University of Pennsylvania and The Children's Hospital of Philadelphia*

In 2007 Medicare introduced Hospital Compare for the public reporting of AMI mortality, providing the public with a predicted mortality rate P over expected "P/E" where P was estimated using a hierarchical model or Bayesian approach based only on patient characteristics. The P/E approach was adopted in order to reduce the noise associated with reporting observed over expected "O/E" in low volume hospitals. This talk will explore how the decision to exclude hospital characteristics such as volume in the model to estimate P gives the false impression that small hospitals with low volume have average mortality rates, when as a group they have mortality rates far above average, because small hospital P/E results are being shrunk to the mean of all hospitals. It will be shown that: (1) in AMI there is a very important volume-outcome relationship; and (2) the hierarchical model without volume as a predictor eradicates the volume-outcome relationship, thereby under-stating mortality at small hospitals; (3) that including volume in the hierarchical model reduces if not eliminates this error. Finally, (4) the talk will discuss the challenges in implementing a hierarchical model with or without hospital characteristics for use in either public reporting or payment policy.

email: silberj@wharton.upenn.edu

DISCUSSION: STATISTICIANS & HEALTH POLICY

Sharon-Lise T. Normand, Harvard Medical School and Harvard School of Public Health*

I will serve as the discussant of this session, focusing on the science of the presentation, but also from the broader context as the (bio)statistician's role (or lack of) in health policy.

email: sharon@hcp.med.harvard.edu

14. STATISTICAL EVALUATION OF DIAGNOSTIC PERFORMANCE USING ROC ANALYSIS

COMBINING BIOMARKERS TO IMPROVE DIAGNOSTIC ACCURACY

Chunling Liu, Hong Kong Polytechnic University
Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*
Susan Halabi, Duke University

Diagnostic accuracy can be improved considerably by combining multiple biomarkers. Although the likelihood ratio provides optimal solution to combination of biomarkers, the method is sensitive to distributional assumptions which are often difficult to justify. Alternatively simple linear combinations can be considered whose empirical solution may encounter intensive computation when the number of biomarkers is relatively large. Moreover, the optimal linear combinations derived under multivariate normality may suffer substantial loss of efficiency if the distributions are apart from normality. In this paper, we propose a new approach that linearly combines the minimum and maximum values of the biomarkers. Such combination only involves searching for a single combination coefficient that maximizes the area under the receiver operating characteristic (ROC) curves and is thus computation-effective. Simulation results show that the min/max combination may yield larger partial or full area under the ROC curves and is more robust against distributional assumptions. The methods are illustrated using the growth-related hormones data from the Growth and Maturation in Children with Autism or Autistic Spectrum Disorder Study (Autism/ASD Study).

email: liua@mail.nih.gov

PERFORMANCE EVALUATION IN TASKS OF DETECTION AND LOCALIZATION OF MULTIPLE TARGETS PER SUBJECT

Andriy I. Bandos, University of Pittsburgh*

In many fields, diagnostic tasks frequently require detection and correct localization of possibly multiple targets within a subject. In medical imaging, performance characteristics in such "detection and localization" tasks (e.g., identification of nodules in the lungs, masses in the breast) play an important role at many stages

of development, optimization, regulatory approval, and general acceptance of diagnostic technologies and practices. Diagnostic assessment of a subject in practice often results in a-priori unknown number of findings. Hence, many practically relevant types of performance assessment methods evaluate diagnostic systems under a protocol where the number and locations of the system's responses is unrestricted. The resulting type of data requires nontrivial extensions of the classical ROC analysis (e.g., FROC, ROI, AFROC). The statistical analyses of performance in "detection and localization" tasks have been developing for many years and include various, often interrelated, methods. Common practically relevant issues for all types of analyses include choice of the summary index, its interpretation, and statistical analysis in the presence of complexly structured data. This talk will outline the general problem of performance assessment in detection and localization tasks, major analytical approaches and their relative merits, and will present several recent developments in the field.

email: anb61@pitt.edu

THE USE OF THE INVARIANCE PROPERTY IN ROC ANALYSIS

Kelly H. Zou, Pfizer Inc.*

The receiver operating characteristic (ROC) analysis is a useful tool for assessing the performance of medical tests and for evaluating predictive algorithms. A special feature that an ROC curve is invariant to any monotone transformation is utilized in this research. That is, a single diagnostic test is evaluated on random samples of subjects of two underlying populations governed by the binary gold standard. Empirical methods for fully estimating an ROC curve nonparametrically based on the ranks of the two-sample measurements in a combined ranking are described. Furthermore, it is assumed that transformed test results via a normality transformation of the two samples follow a particular parametric model, namely the binormal model. Several nonparametric and parametric adjustment methods for the effect due to stratification are presented. These methods are compared via Monte-Carlo simulation studies and are exemplified using published data.

email: Kelly.Zou@pfizer.com

15. STATISTICAL APPLICATIONS IN FOOD SAFETY

DISPROPORTIONALITY ANALYSES FOR DETECTION FOOD ADVERSE EVENTS

Stuart J. Chirtel, U.S. Food and Drug Administration*

CFSAN receives spontaneous adverse event reports possibly related to foods, dietary supplements and cosmetics from physicians, individuals, industry and other sources. These data are analyzed for suspicious or unexpected reporting patterns using a variety of techniques including several types of disproportionality analyses. The goal of the analysis is to detect higher-than-expected numbers of reports of product-symptom combinations. Due to the lack of information on product usage and reporting, signal detection techniques rely on deviations in the symptom profile of a specific product compared to that of the database as a whole. The addition of sales data to the analysis of spontaneously reported reports can drastically improve detection power.

email: stuart.chirtel@fda.hhs.gov

STATISTICAL METHODS FOR ANALYSIS OF DNA APTAMERS

Yan Luo, U.S. Food and Drug Administration*
Jeffrey A. DeGrasse, U.S. Food and Drug Administration
Sara Handy, U.S. Food and Drug Administration
Andrea Ottesen, U.S. Food and Drug Administration
Errol Strain, U.S. Food and Drug Administration

There is a continual need for rapid, robust, and field deployable assays to detect and quantify adulterants in food products. While antibodies are extensively used in these assays, the use of aptamers, which are nucleic acids that specifically bind to a target molecule, is currently under evaluation at the FDA's Center for Food Safety and Applied Nutrition (CFSAN). Aptamers are discovered by a process known as Systematic Evolution of Ligands by Exponential Enrichment (SELEX), a well established in vitro selection tool to analyze the sequences with high affinity binding to a ligand. SELEX begins with a random population (~6 x 1014 unique sequences) of ssDNA and, through a series of selection rounds, progressively enriches the sequence space by selecting those molecules with the desired binding activity. In this work, the sequence space from different enrichment rounds are characterized using next generation sequencing techniques. This data allows both for an observation of the evolving sequence space as it progresses through selections rounds, and also the determination of top aptamer candidates. This study examines the utility of DNA microarray statistical methods for analysis of aptamer sequence data from the SELEX experiment.

email: yan.luo@fda.hhs.gov

FORENSIC ANALYSIS OF BACTERIAL GENOMES FOR FOODBORNE OUTBREAKS

Errol A. Strain, U.S. Food and Drug Administration*
Allard Marc, U.S. Food and Drug Administration
Eric Brown, U.S. Food and Drug Administration
Luo Yan, U.S. Food and Drug Administration

The FDA is using Next Generation DNA Sequencing (NGS) techniques to produce whole genome shotgun sequences of bacteria related to foodborne outbreaks. The goal of these types of analyses is to find links between food, environmental, and clinical bacterial isolates collected as part of the investigation. While there are some similarities between the genetic analysis of human and bacterial DNA, the clonal, haploid nature of the bacteria and small sequence databases confound forensic interpretation. The study will present the forensic analysis of a Salmonella Enteritidis outbreak associated with eggs.

email: Errol.Strain@fda.hhs.gov

CONFIDENCE INTERVALS FOR COUNTING MICROBES ON PLATES

Robert Blodgett, PhD (FDA/CFSAN)

The quantification of microorganisms is often performed by growing the bacteria on culture plates. Counts of bacteria, or colonies, on the plates of microbes may include some plates that are too numerous to count (TNTC). Estimates of the concentration of target microbes including these TNTC plates can use methods from maximum likelihood or imputation. Both of these methods and a normal approximation can produce confidence intervals which are compared.

16. SYNTHETIC HEALTH DATA FOR CONFIDENTIALITY CONTROL

IMPUTATION OF CONFIDENTIAL DATASETS WITH SPATIAL LOCATIONS USING POINT PROCESS MODELS

Thais V. Paiva, Duke University*
Jerome P. Reiter, Duke University

We suggest a method to generate synthetic data sets with imputed spatial locations, respecting individuals' confidentiality without losing statistical utility. Our primary interest is to impute the spatial coordinates conditional on the response and explanatory variables. We generate the imputed data sets using spatial point models. The underlying spatial intensities are modeled allowing flexible relationships among the variables and the spatial locations. Using a Bayesian framework, we obtain posterior samples of the intensities, and use them to generate imputed data sets for public release. We verify the quality of the synthetic data, along with the level of confidentiality.

email: tvp@stat.duke.edu

MULTIPLE IMPUTATION USING CHAINED EQUATIONS FOR HIGH DIMENSIONAL LONGITUDINAL MISSING DATA IN THE DCCT/EDIC STUDY

Michael D. Larsen, The George Washington University*
Paula McGee, The George Washington University
John M. Lachin, The George Washington University

The Diabetes Control and Complications Trial (DCCT; 1983-93) compared the effect of conventional (n=729) versus intensive (n=711) treatment for type 1 diabetes on diabetes complications. The Epidemiology of Diabetes Interventions and Complications (EDIC) is an observational follow-up study of the DCCT cohort. The demonstrated long term benefits of intensive therapy on diabetes complications have established it as the standard of care. During the DCCT blood glucose profiles (7 measurements in a day) were collected every 3 months over 9 years. Extensive missing data has compromised the ability to evaluate the association of diurnal patterns of glucose levels and variation with long term diabetic complications. Of 95,375 expected glucose measurements within the profiles, only 79,499 (83%) were obtained. Multiple imputation using chained equations (mice) is a realistic approach to filling in the missing values and enabling specific analyses. More than 250 possible glucose values per subject and the high fraction of missing values and their autocorrelations present a variety of challenges. Selection of models and predictors along with assessment of imputation quality are presented. Relevance of this work for synthetic high dimensional data for disclosure control will be discussed.

email: mlarsen@bsc.gwu.edu

ASSESSING THE PRIVACY OF RANDOMIZED MULTIVARIATE QUERIES TO A DATABASE USING THE AREA UNDER THE RECEIVER-OPERATOR CHARACTERISTIC CURVE

Gregory J. Matthews, University of Massachusetts*
Ofer Harel, University of Connecticut

As the amount of data generated continues to increase, consideration of individuals' privacy is a growing concern. As a result, there has been a vast quantity of research done on methods of statistical disclosure control (SDC). Some of these methods propose to release a randomized version of the data rather than the actual data. While methods of this type certainly offer some layer of protection since no actual data is released, there is still the potential for private information to be disclosed. Quantifying the level of privacy provided by these methods is often difficult. In the past, a method for assessing privacy using the receiver-operator characteristic (ROC) curve based on ideas related to differential privacy has been proposed. However, the method was only demonstrated for univariate randomized releases. Here, the ROC based privacy measure is extended to randomized multivariate queries.

email: gjm112@gmail.com

DISCLOSURE CONTROL IN THE CanCORS

Bronwyn Loong, Harvard University*
David Harrington, Harvard University
Alan Zaslavsky, Harvard Medical School
Yulei He, Harvard Medical School

The Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium is a multisite, multimode, multiwave study of the quality and patterns of care delivered to population and health system based cohorts of newly diagnosed patients with lung and colorectal cancer. The Consortium is committed to sharing the data gathered during its work to the widest possible audience without comprising the confidentiality of respondents' identities and sensitive attributes. To satisfy these requirements, the consortium can release partially synthetic data, where values of high disclosure risk variables for the units originally surveyed are replaced with multiple imputations. In this talk, we discuss partial synthesis of the CanCORS patient survey lung cancer data set, focusing on key decision steps in selecting variables for synthesis, selection of imputation models and measurements of data utility and disclosure risk. We use the sequential regression multiple imputation method to generate the synthetic data and use stepwise regression to select predictors for each variable to be imputed. Data utility is evaluated by comparison of original data and partially synthetic data analytic results, based on statistical models motivated by two published analyses on the original data. Our work illustrates the partial synthesis of a large-scale multiobjective health survey.

email: bloong@fas.harvard.edu

PARTIAL SYNTHETIC DATA FOR POPULATION-BASED CANCER REGISTRY DATA

Mandi Yu, National Cancer Institute, National Institutes of Health*
Li Zhu, National Cancer Institute, National Institutes of Health
Benmei Liu, National Cancer Institute, National Institutes of Health
Eric (Rocky) Feuer, National Cancer Institute, National Institutes of Health
Kathleen Cronin, National Cancer Institute, National Institutes of Health

The NCI's SEER Program collects and publishes data about cancer patient demographics, geographic locations, tumor and treatment information from population-based cancer registries. It has been the most authoritative source of data for describing cancer incidence and survival. Increasingly, researchers are demanding the access to small area data to identify areas with elevated cancer rates and to plan and monitor the impact of cancer control and prevention activities at local levels. However, confidentiality breach is likely when one combines detailed geography with basic demographics. The authors developed a multiple imputed partial synthetic data approach to alter certain demographic data without distorting the statistical information in a cancer registry microdata at local levels. The data were based on patients who were diagnosed in 2008 from one of three registries in California. The spatial structure on county level demographics was captured through a Bayesian hierarchical model. This approach does not release any actual patients' demographic data, thus reduces the risk of re-identification. The analytic validity was evaluated by comparing cancer incidence estimated from the synthetic data with those obtained from the original data.

email: yum3@mail.nih.gov

17. STATISTICAL ISSUES ARISING FROM ALTERNATIVES TO DOUBLE-MASKED RANDOMIZED CONTROLLED TRIALS

A REGULATORY VIEW OF THE STATISTICAL CHALLENGES FOR ALTERNATIVES TO DOUBLE-MASKED RANDOMIZED CONTROLLED TRIALS

Gregory Campbell, U.S. Food and Drug Administration*

While the double-masked, randomized controlled trial is often the ideal design for investigating the safety and effectiveness of most new medical products, there are often reasons which make such a design impossible or highly impractical. Deviations from this ideal can pose problems in terms of assessing the size of the bias or in analyzing the data. The issues can run the gamut from deciding whether the mask may have been broken, what to do if the control is concurrent but non-randomized, the use of a historical control, to uncontrolled one-arm studies without controls that rely instead on Objective Performance Criteria (OPCs) or on other performance goals. The unique challenges for diagnostic devices will also be discussed. These issues will be discussed from both a regulatory and a statistical perspective.

email: greg.campbell@fda.hhs.gov

STUDY DESIGN AND ANALYSIS ISSUES WITH EFM-CAD

Bipasa Biswas, U.S. Food and Drug Administration*

Evaluations of diagnostic tests are based on performance measures which do not usually involve a randomized double-arm clinical trial. In a study comparing a new device to another comparator device the subjects usually get both tests. The safety and effectiveness of a diagnostic test is usually inferred from adequate diagnostic performance that has been well characterized. A particular type of diagnostic device, Electronic fetal monitors, can be classified as antepartum (before labor) or intrapartum (during labor). This talk focuses on intrapartum fetal monitors which are classified into three types- type I providing basic heart rate patterns, type II which detects specific patterns and type III which stratifies to various risk categories of future events of clinical concerns. Various study designs for evaluating type III and some type II devices will be discussed.

email: bipasa.biswas@fda.hhs.gov

ASSESSING THE "SUCCESS" OF THE BLIND IN SHAM-CONTROLLED RANDOMIZED CLINICAL TRIALS

Valerie Durkalski, Medical University of South Carolina*
Qi Wu, Medical University of South Carolina

A critical component to randomized-controlled clinical trials is the inclusion of adequate treatment blinding to help ensure unbiased estimates of treatment effects. Although a common design feature, several trials, particularly device and surgical trials, are challenged to develop adequate controls. When feasible, these trials attempt to preserve the blind by developing a "sham" control that mimics the experimental treatment. In these cases, it is important to assess the quality of blinding and the impact on the treatment estimates. Options for assessment include questionnaires of "best" treatment guess and confidence in the guess which should be collected at multiple timepoints throughout the trial period. We examine the use of these questionnaires in sham-control trials and the relationship between blinding quality and treatment effect.

email: durkalsv@musc.edu

PRACTICES OF USING PROPENSITY SCORE METHODS IN DRUG-ELUTING STENT STUDIES

Hong Wang, Boston Scientific Corporation*
H. Terry Liao, Boston Scientific Corporation

The propensity score methods have been used primarily to reduce bias and increase precision in observational studies. Specifically the matching methodology is most applied to drug eluting stent studies to assess the clinical outcomes with balanced baseline characteristics for data integration of single-arm studies or non-randomized trials. The propensity-adjusted clinical outcomes may be used for planning of future clinical trials, post-hoc analyses for specific subgroups, and justification of business strategy. The matching models and algorithms may vary due to the size of a match and the number of treatments. The most frequent seen scenario is one-to-one match in a two-treatment comparison. A real-world case will be presented. In addition to the most common model, one example for one-to-many match and one application for three-treatment will be presented for discussion.

email: Hong.Wang@bsci.com

STUDY DESIGNS FOR POSTMARKET SURVEILLANCE

Theodore Lystig, Medtronic, Inc.*
Jeremy Strief, Medtronic, Inc.

Postmarket surveillance studies for medical devices may be mandated by the FDA under Section 522 of the Federal Food, Drug and Cosmetic Act. The general and specific content of a postmarket surveillance plan is described both in 21 CFR 822 and in the recently issued draft guidance document on postmarket surveillance. The design of such studies poses special problems, not least of which is that clinical equipoise may no longer exist between a newly approved device and the control device used in the study which was the basis for marketing approval. Not only is it problematic to conduct a randomized trial, but it may even be more defensible to use a single arm observational study design. This talk will describe when it is appropriate to use single arm observational studies in postmarket surveillance, including why it may be desired to motivate sample size more from an estimation than from a testing standpoint.

email: theodore.lystig@medtronic.com

CHALLENGES IN NON-RANDOMIZED CONTROLLED MEDICAL DEVICE TRIALS

Shelby Li, Medtronic, Inc.*
Shufeng Liu, Medtronic, Inc.

Double-blinded randomized controlled clinical trial design is commonly used to test safety and efficacy of a medical treatment/device therapy. Often, blinding or randomization, or neither is practical to study a cardiac device. Alternative designs can be: randomization without blinding, treatment-control allocation without randomization, or one-armed Objective Performance Criteria evaluation, etc. A few recent conducted clinical medical device studies will be discussed to evaluate the pros and cons of these alternative study designs. Conclusion: randomization is a key for comparing two treatment groups. Statistical methods, such as covariates adjustments or propensity score analysis should be employed for observational case-control studies.

email: shelby.li@medtronic.com

18. STATISTICAL GENETICS

NONLINEAR SUFFICIENT DIMENSION REDUCTION FOR ASSOCIATION TESTING OF COMPLEX TRAITS

Hongjie Zhu, Duke University*
Lexin Li, North Carolina State University
Hua Zhou, North Carolina State University

Association tests based on next-generation sequencing data are often under-powered due to presence of rare variants and large amount of neutral or protective variants. A successful strategy is to aggregate genetic information within meaningful SNP-sets, e.g., genes or pathways, and test association on SNP-sets. Many existing methods for group-wise tests require specific assumptions about the direction of individual SNP effects and/or perform poorly in the presence of interactions. We proposed a

joint association test strategy based on kernel sufficient dimension reduction methods to meet the challenges. Accompanying this strategy, we also propose a class of new kernels specially designed for genotype data, which can potentially boost the power of various kernel-based methods. The strategy coupled with the new kernels shows superior performance over existing methods over various disease models simulated from sequence data of real genes.

e-mail: hongjie.zhu@hotmail.com

LOCAL ANCESTRY INFERENCE IN ADMIXED NUCLEAR FAMILIES USING A TWO-LAYER CLUSTERING MODEL

Wei Chen, University of Pittsburgh*
Yongtao Guan, Baylor College of Medicine

In admixed populations such as Hispanics and African American, the genome of each individual consists of chromosome segments from two or more populations. Accurate global and local ancestry estimations with dense genetic markers are crucial to disease association analyses and demographic history inference. To date, there is no existing method for local ancestry inference in families. We proposed a computationally efficient two-layer clustering model approximating the classical coalescent tree by a class of sub-trees. The top layer models different ancestry populations and the bottom layer models the subtle differences among observed haplotypes based on Linkage-disequilibrium (LD) information. Families were jointly modeled to represent the inheritance constraints. Through simulation, we compared our method with the widely used software HAPMIX and LAMP for unrelated individuals. We show that our method greatly increases the global and local ancestry inference when family members are available and taken into account. We applied our method to a study of chronic obstructive pulmonary disease (COPD) families in Costa Rica.

e-mail: weichen.mich@gmail.com

A FLEXIBLE VARYING COEFFICIENT MODEL FOR THE DETECTION OF NONLINEAR GENE-ENVIRONMENT INTERACTION

Yuehua Cui, Michigan State University
Cen Wu, Michigan State University*

The genetic influences on complex disease traits are generally dependent on the joint effects of disease variants, environment factors, as well as their interplays. Gene-environment (G×E) interactions play vital roles in determining an individual disease risk, but the underlying machinery is poorly understood. The linear assumption for the relationship between genetic and environmental factors, along with their interactions, prevails in the current regression-based framework to examine the G×E interaction. This assumption, however, could be violated by the nature of non-linear G×E interaction. As an extension to our previous work on continuous traits, we propose a flexible varying coefficient model for the detection of nonlinear G×E interaction for binary traits. The varying coefficients are modeled by a non-parametric regression function through which one can model the nonlinearity of G×E interaction. A group of statistical tests is proposed to elucidate the machinery of G×E interaction. The utility of the proposed method is illustrated via simulation and real data analysis.

e-mail: wucen@stt.msu.edu

PERMUTATION-BASED EXPRESSION PATHWAY ANALYSIS, WITHOUT PERMUTATION

Yi-Hui Zhou, University of North Carolina at Chapel Hill*
Fred A. Wright, University of North Carolina at Chapel Hill

Resampling-based expression pathway analysis techniques have been shown to preserve type I error, in contrast to simple gene-list approaches which implicitly assume independence of genes in ranked lists. However, resampling is intensive in computation time and memory requirements. We describe highly accurate analytic approximations to permutations of score statistics, including novel approaches for Pearson correlation and summed score statistics, that have good performance for even relatively small sample sizes. In addition, the approach provides insight into the permutation approach itself, and summary properties of the data that largely determine the behavior of the statistics. Within the framework of the SAFE pathway analysis procedure, our approach preserves the essence of permutation analysis, but with greatly reduced computation. Extensions to include covariates are described, and we test the performance of our procedures using simulations based on real datasets of modest size.

e-mail: yihui2006@gmail.com

RESPONSE-SELECTIVE SAMPLING DESIGNS FOR RARE VARIANT ANALYSIS IN GENETIC ASSOCIATION STUDIES

Yildiz E. Yilmaz, University of Toronto*

The cost of exome or whole genome sequencing of an entire GWAS cohort to test for association with rare variants is prohibitive, and selection of individuals for sequencing according to their quantitative trait (QT) value can improve cost-efficiency. We examine QT-dependent sampling designs, including an extreme phenotype design and an inverse probability selection (IPS) design in which all individuals have a non-zero probability of being selected into the sample, but those with extreme phenotypes have a proportionately higher probability. We apply methods for

two-phase sampling designs that use semiparametric maximum likelihood estimation to fit a regression model for association of the QT with a rare variant score. For the IPS design, we compare maximum likelihood with inverse probability weighting of estimating equations. We investigate the effect of assigned individual selection weights on properties of parameter estimates and the power of the association test. In evaluations of sampling designs and methods by simulation, we found that QT-dependent sampling designs are generally more efficient than a simple random sample of the same number of individuals. For most designs examined, semiparametric maximum likelihood provides efficient estimation and more powerful tests than the inverse probability weighting approach.

e-mail: yilmaz@lunenfeld.ca

19. SPATIAL/TEMPORAL MODELING

A GEOADDITIVE IMPUTATION APPROACH TO MEASUREMENT ERROR CORRECTION WITH SPATIALLY MISALIGNED NON-NORMAL DATA

Lauren Hund, Harvard School of Public Health*
Till Baernighausen, Harvard School of Public Health
Frank Tanser, Africa Centre for Health and Population Studies
Brent Coull, Harvard School of Public Health

In a longitudinal observational study in rural South Africa, study investigators aim to quantify the association between HIV incidence and latent non-normal spatial covariates (e.g. average ART coverage, predicted from individual ART usage). The exposure locations are different from the outcome locations, resulting in spatial point-to-point misalignment. In this setting, existing frequentist methods (plug-in estimators) for examining the association between an exposure and outcome are inefficient and potentially biased, as a result of increased classical measurement error in the non-normal setting. We propose an imputation procedure for predicting latent spatial covariates; by jointly modeling the exposure and outcome, we increase the efficiency of the estimator and reduce bias relative to the plug-in estimators. To speed up the computation time in the joint exposure/outcome model, we introduce a fast multivariate spatial regression model which naturally handles all types of spatial misalignment (point-to-point, block-point, and block-to-block misalignment).

email: lbhund@gmail.com

MODELING AIR POLLUTION MIXTURES IN SOUTHERN CALIFORNIA

Reza Hosseini, University of Southern California
Meredith Franklin, University of Southern California*
Duncan Thomas, University of Southern California
Kiros Berhane, University of Southern California

Air quality has consistently been shown to be an important determinant of public health. However, characterizing the complex behavior of the multiple pollutant mixture inherent in ambient air is a challenging statistical problem. This work develops a multivariate model that accounts for correlation between pollutants across time and space by employing latent spatial processes. A Bayesian hierarchical approach is used to estimate the parameters and predict pollution levels at unobserved times and locations. To obtain the posterior distributions, the unnormalized posterior is calculated on a grid of the parameters. In order to make the computations feasible parallel computing is employed to calculate the unnormalized posterior on several partitions of the grid. Application of the model incorporates air pollution concentrations (NO2, NOx and O3) gathered at several locations within 12 Southern California communities during three periods in 2005-2006. Several traffic-related covariates are also included to improve the predictions.

email: meredith.franklin@usc.edu

FLEXIBLE BAYESIAN PREDICTIVE PROCESS SPATIAL FACTOR MODELS FOR MISALIGNED DATA SETS

Qian Ren, University of Minnesota*
Sudipto Banerjee, University of Minnesota

Spatial factor analysis model is used to separate the sources of variation according to the spatial scales and reduce the model dimension. In the spatial context, dimension reduction is also required with respect to the number of observed locations. Here, we demonstrate how a dimension-reducing low-rank spatial process (called a predictive process) leads to a class of computationally feasible spatial factor analysis model, thereby reducing the computational burden. We also address the important practical problem of how to select the random component in the hierarchical model. The latent spatial processes are allowed to effectively drop out of the model by using indicator variables. A Markov chain Monte Carlo (MCMC) algorithm was developed for estimation with an emphasis toward missing data. The missing data problem in spatial factor analytic settings is complicated by the spatial misalignment of outcomes. This pertains to the rather commonplace settings, where all the outcomes have not been observed over a common set of locations. We present sampling-based methods that condition on the observed data and recover the full posterior distribution of the missing values in a Bayesian predictive framework. We illustrate our methodology with simulated data and an environmental data set.

email: renxx014@umn.edu

HIGH-DIMENSIONAL STATE SPACE MODELS FOR DYNAMIC GENE REGULATORY NETWORKS

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

Gene regulation has been extensively studied on many levels to describe biological systems. Expression profiles from time-course experiments along with appropriate models will allow us to identify dynamic regulatory networks on the fundamental level. The challenges fall on the high-dimensional nature of such data without a surprise. We propose a high-dimensional linear State Space Model (SSM) with a new Expectation-Regularized-Maximization (ERM) algorithm to construct the dynamic gene regulatory network. System noise and measurement error can be separately specified through SSMs. However, a high-dimensional SSM gives us too many parameters to estimate. The proposed new ERM algorithm uses the idea of the adaptive Lasso-based variable selection method so that the sparsity property of gene regulatory networks can be preserved. The proposed method is applied to identify the dynamic GRN for yeast cell cycle progression data.

email: sinuiris@gmail.com

A STOCHASTIC AND STATE SPACE MIXTURE MODEL OF HUMAN LIVER CANCER MULTIPLE-PATHWAY MODEL INVOLVING BOTH HEREDITARY AND NON-HEREDITARY CANCER

Xiaowei (Sherry) Yan, Geisinger Center for Health Research*
Wai-Yuan Tan, University of Memphis

Based on recent biological studies, we have developed a state space mixture model for human liver cancer. The state space model joins stochastic system model with a statistical model, in which the stochastic system model composes of two parts: first is a stochastic model involving 2 different pathways for Non-hereditary liver cancer, the second part is hereditary pattern, which was represented by a mixture model. To this end, the probability of liver cancer developed from each pathway was derived. Then the statistical model combines the liver cancer incidence rate with observational data. Based on this model we have developed a generalized Bayesian approach to estimate the parameters through the posterior modes of the parameters via Gibbs sampling procedures. We have applied this model to fit and analyze the SEER data of human liver cancer incidence from NCI/NIH. Our results indicate that the model not only provides a logical avenue to incorporate biological information but also fits the data much better than other models including the 4-stage single pathway model. This model not only would provide more insights into human liver cancer but also would provide useful guidance for its prevention and control and for prediction of future cancer cases.

email: xwyan2001@yahoo.com

20. NON-LINEAR, PK-PD, AND DOSE-RESPONSE MODELS

NON-LINEAR MODELS FOR MULTIPLE FLOW EXHALED NITRIC OXIDE DATA

Sandrah P. Eckel, University of Southern California*
Kiros Berhane, University of Southern California
William S. Linn, University of Southern California
Muhammad T. Salam, University of Southern California
Yue Zhang, University of Southern California
Edward B. Rappaport, University of Southern California
Frank D. Gilliland, University of Southern California

The fractional concentration of exhaled nitric oxide (FeNO) is thought to be a marker for airway inflammation and has been associated with air pollution exposure. A deterministic non-linear two-compartment model describes the physiology of NO production in the respiratory system. Regression models approximating the two-compartment model can be estimated using a small number of repeated FeNO measurements at multiple exhalation flow rates. The coefficients are interpreted as parameters governing NO production in different anatomical locations and may provide insight into the mechanisms of airway inflammation, particularly related to air pollution exposure. Multiple flow data originated in small experimental studies, but is now available in a large cohort of children in the Southern California Children's Health Study (CHS) which also has extensive data on air pollution exposures. Methods for
effectively modeling such data have not been developed.
We develop and evaluate methods to: estimate physiologic parameters, either separately for each participant or by pooling across participants using a mixed-effect model (Stage I) and relate estimated parameters to environmental exposures (Stage II). Methodological challenges include producing parameter estimates within a biologically plausible range of values and non-linear mixed-effect models with non-normal random effects.

email: eckel@usc.edu

AN EMPIRICAL APPROACH TO SUFFICIENT SIMILARITY: COMBINING EXPOSURE DATA AND MIXTURES TOXICOLOGY DATA

Chris Gennings, Virginia Commonwealth University*
Scott Marshall, BioStat Solutions, Inc.
LeAnna G. Stork, Monsanto Company

U.S. EPA guidance documents for cumulative risk assessment of chemical mixtures describe a whole mixture approach where mixture-specific toxicity data considered sufficiently similar to mixtures in the environment may be used as a surrogate; however, specific information on how to define sufficiently similar mixtures is not provided. Herein, we define sufficient similarity, without assuming additivity, using equivalence testing methodology comparing the distance between benchmark dose estimates for mixtures in both data rich and data poor cases. We use a "mixtures Hazard Index" on sufficiently similar mixtures linking exposure data with mixtures toxicology data. The methods

are illustrated using pyrethroid mixtures data reported in floor wipe samples collected in a nationally representative survey of child care centers and dose-response data for the acute effects of mixtures of pyrethroids from laboratory animal studies on neurobehavioral function. Accounting for the relative potency of the chemicals, the mixtures from 90% of the centers where at least one pyrethroid was detected (75% of centers) were determined to be sufficiently similar. The approach offers an alternative strategy for risk evaluation of typical mixtures that bypasses the general assumption of dose additivity (Supported by #R01ES015276, #UL1RR031990, #T32ES007334).

email: gennings@vcu.edu

COMPARISON OF DIFFERENT BIOSIMILARITY CRITERIA UNDER VARIOUS STUDY DESIGNS

Eric Chi, Amgen Inc.
Shein-Chung Chow, Duke University
Hao Zhang, Amgen Inc.*

For the assessment of biosimilarity of follow-on-biologics, the classical approach for assessing bioequivalence (for small molecule drug products) may not be appropriate due to fundamental differences between the small molecule drug products and biological drug products. For example, biological drug products are not only more variable, but also more sensitive to small changes during the manufacturing process. As a result, criteria similar to the assessment of population/individual bioequivalence and a scaled average biosimilarity criterion are suggested. In this presentation, assessment of biosimilarity of follow-on-biologics based on different biosimilarity criteria under various study designs is examined. The purpose of this research is not only to compare these criteria under various study designs, but also to select the most appropriate design for a certain criterion. Furthermore, the relationships among these criteria and the condition under which they are equivalent or related are examined either theoretically or through simulations.

email: haoz@amgen.com

SEMIPARAMETRIC MODELING OF DOSE-RESPONSE RELATIONSHIPS IN EX-VIVO EXPERIMENTS

Samiha Sarwat, Indiana University School of Medicine*
Jaroslaw Harezlak, Indiana University School of Medicine
Clarissa Valim, Harvard School of Public Health

In medical studies, dose-response relationship may describe changes in an organism caused by different drug doses after a fixed time period. Dose-response data are often collected in drug assays used to monitor the development of resistance. Data points obtained from a biological sample are often highly correlated, because their aliquots are subjected to varying drug doses and more than one replication of the sample is assayed (technical replicates). The analysis of such experiments frequently relies on a parametric model, generally a sigmoidal (logistic) function. However, when the dose-response curve does not follow a parametric function, non-parametric methods are necessary. We propose an extension to a penalized regression spline semiparametric model (Ruppert et.al., 2003) that allows modeling of the smooth dose-response relationships with correlated data via the linear mixed model representation. The proposed method preserves the hierarchy of the technical and biological replicates while letting the data guide the mean model estimates. The quantities of interest, for example IC50, are obtained and their properties are derived. We illustrate the method on simulated data and apply it to analyze ex vivo drug assays in malaria monitoring drug resistance.

email: harezlak@iupui.edu

NONLINEAR MODELS FOR META-ANALYSIS OF SUMMARY EXPOSURE-RESPONSE DATA

Paul W. Stewart, University of North Carolina at Chapel Hill*
Vernon Benignus, U. S. Environmental Protection Agency

This report focuses on a methodological challenge encountered in using nonlinear mixed-effects models for analysis of continuous exposure-response data from human subjects. Specifically, investigations of responses to toxic environmental exposures must sometimes rely on historical pooled summary data from a number of previously published studies in lieu of conducting controlled, randomized dose-response experiments. The summary data take the form of dose-specific average responses calculated by averaging across individuals within each study. In this instance the original raw data were not available and were unobtainable. This limitation poses two challenges: the available average values are subject to variance heterogeneity due to the varying sample sizes of the published studies, and appropriate dose-response models for the averaged values are ill-suited to inference if individual-specific nonlinear random effects are important sources of variance. We discuss strategies for specifying and fitting statistical models which account for the varying sample sizes and cope with individual-specific nonlinear random effects. Analysis of environmental toluene dose-response data from human subjects is illustrated with attention to estimation of dose-response curves and confidence limits.

email: paul_stewart@unc.edu

STATISTICAL INFERENCE FOR DYNAMIC SYSTEMS GOVERNED BY DIFFERENTIAL EQUATIONS WITH APPLICATIONS TO TOXICOLOGY

Siddhartha Mandal, University of North Carolina at Chapel Hill*
Pranab K. Sen, University of North Carolina at Chapel Hill
Shyamal D. Peddada, National Institute of Environmental Health Sciences, National Institutes of Health

Deterministic and stochastic differential equations are commonly used to describe a wide variety of biological and physiological phenomena. For example, they are used in physiologically based pharmacokinetic (PBPK) models for describing absorption, distribution, metabolism and excretion (ADME) of a chemical in animals. Parameters of PBPK models are important for understanding the mechanism of action of a chemical and are often estimated using iterative non-linear least squares methodology. However, one of the challenges with the existing methodology is that one cannot readily obtain the uncertainty estimates associated with the parameter estimates. Secondly, the existing methodology does not account for variability between and within animals. Using functional data analytic methodology, in this article we develop a general framework for drawing inferences on parameters in models described by a system of differential equations. The proposed methodology takes into account variability between and within animals. The performance of the proposed methodology is evaluated using a simulation study mimicking a real data set and the methodology is illustrated using a data obtained from a benzene inhalation study.

email: sid.stat.iitk@gmail.com

A B-SPLINE BASED SEMIPARAMETRIC NONLINEAR MIXED EFFECTS MODEL

Angelo Elmi, George Washington University*
Sarah Ratcliffe, University of Pennsylvania School of Medicine
Samuel Parry, University of Pennsylvania School of Medicine
Wensheng Guo, University of Pennsylvania School of Medicine

The Semiparametric Nonlinear Mixed Effects Model (SNMM) (Ke and Wang, 2001) provides a flexible framework for longitudinal comparisons of curve shapes between groups. In this article, we develop an alternative method for fitting the SNMM by reformulating Ke and Wang's smoothing spline based model in terms of B-splines. The existing algorithm is based on a backfitting procedure that iterates between two mixed models whose corresponding likelihoods are not equivalent to the likelihood of all model parameters. The consequence is a lack of reliable convergence and statistical inference. Using B-splines, however, overcomes these disadvantages by simplifying the likelihood computations without sacrificing model flexibility. Therefore, the algorithm can be expressed in terms of existing, accurate techniques based on Adaptive Gaussian Quadrature. The model is applied to labor curves, cervical dilation measured longitudinally, from women attempting a vaginal birth after cesarean. Only partial curves were measured on cases of uterine rupture given the need for emergency c-section while controls completed delivery naturally. The model allowed us to estimate and compare the average labor curve shape between cases and controls and also determine the earliest time at which clinicians could distinguish between the average labor curves in different groups.

email: sphafe@gwumc.edu

21. LONGITUDINAL DATA

A BAYESIAN SEMIPARAMETRIC APPROACH FOR INCORPORATING LONGITUDINAL INFORMATION ON EXPOSURE HISTORY FOR INFERENCE IN CASE-CONTROL STUDIES

Dhiman Bhadra, Worcester Polytechnic Institute*
Michael J. Daniels, University of Florida
Sung Duk Kim, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Malay Ghosh, University of Florida
Bhramar Mukherjee, University of Michigan

Case-control studies primarily compare the exposure distribution of a group of cases to a group of controls to identify potential risk factors of a disease. In a typical case-control study, exposure information is collected at a single timepoint for the cases and controls. However, case-control studies are often embedded in existing cohort studies containing longitudinal exposure history on the participants. Recent medical studies have indicated that incorporating past exposure history, when available, may lead to more precise estimates of the disease risk. In this paper, we propose a flexible Bayesian semiparametric regression approach to jointly model the time-varying exposure profiles of the cases and controls and also the influence pattern of the exposure profile on the disease status. This enables us to analyze how the present disease status of a subject is influenced by his/her past exposure history conditional on the current ones and properly account for uncertainties associated with both stages of the estimation process in an integrated manner. Analysis is carried out in a hierarchical Bayesian framework using Reversible jump Markov chain Monte Carlo (RJMCMC) algorithms. The proposed methodology is motivated by, and applied to a nested case-control study of prostate cancer where longitudinal biomarker information is available for the cases and controls.

email: dbhadra@wpi.edu

SHARED PARAMETER MODELS FOR LONGITUDINAL MULTIPLE SOURCE COST DATA

Mulugeta Gebregziabher, Medical University of South Carolina and Ralph H. Johnson VA Medical Center, Charleston*
Yumin Zhao, Ralph H. Johnson VA Medical Center, Charleston
Clara E. Dismuke, Ralph H. Johnson VA Medical Center, Charleston
Kelly J. Hunt, Ralph H. Johnson VA Medical Center, Charleston and Medical University of South Carolina
Leonard E. Egede, Ralph H. Johnson VA Medical Center, Charleston

Several approaches including transformation, generalized linear mixed models (GLMM) and semi-parametric two part models are used to account for heteroscedasticity, skewness and zeros in healthcare cost analysis. Analyzing aggregated total cost from separate service sources is another critical problem since it could hide factors that have a differential impact on cost sources and lead to incorrect conclusions due to failure to account for the shared correlation among cost variables. We propose a multivariate GLMM (mGLMM) approach that addresses this problem. We demonstrate mGLMM using data from a national cohort of 892,223 veterans with diabetes (followed 2002-2006) to jointly model cost outcomes from inpatient, outpatient and pharmacy services. The joint modeling approach allows assessment of differential covariates effects on each cost type accounting for shared correlation and relevant covariates. We compare log-normal, gamma and exponential distributions to assess whether the proposed joint modeling is robust to distributional assumptions. Goodness of fit measures scaled to sample size, residual by predicted plot and standard error of estimated parameter are used for model comparison. Our results indicate that ignoring correlation among multivariate outcomes could lead to erroneous conclusions as well as biased estimates of cost projections.

email: g.eastham.gilbert@gmail.com

A MIXTURE OF MARKOV MODELS FOR HETEROGENEOUS LONGITUDINAL ORDINAL DATA WITH APPLICATIONS TO ANALYZING LONGITUDINAL BACTERIAL VAGINOSIS DATA

Kyeongmi Cheon, 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
Marie Thoma, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Bacterial vaginosis (BV), characterized by disturbances in vaginal flora, is a recurrent condition that has been implicated in adverse pregnancy outcomes such as preterm birth. While longitudinal ordinal severity measurements have typically been analyzed using Markov models, we propose a new mixture model that flexibly incorporates heterogeneity in the transition process across individuals. We consider a transition process with three severity levels (low, medium, and high). Reflecting features in longitudinal BV data, we consider a mixture of three Markov processes: processes which (1) transitions between the medium and high severity states, (2) transitions between the low and medium severity states, and (3) transition between all three states. This mixture of Markov processes extends the mover stayer model introduced by Blumen, Kogan, and McCarthy (1955), to ordinal longitudinal data where a stayer is newly defined as transitioning between two states (either processes (1) or (2)). We incorporate covariates into each of the three processes as well as in the probability of being in each of the three groups. We use the model to examine the role of HIV infection on the natural history of BV using data from a longitudinal BV cohort.

email: katie.cheon@gmail.com

SEMIPARAMETRIC REGRESSION WITH NESTED REPEATED MEASURES DATA

Rhonda D. VanDyke, Cincinnati Children's Hospital Medical Center*
Resmi Gupta, Cincinnati Children's Hospital Medical Center
Raouf S. Amin, Cincinnati Children's Hospital Medical Center

Semiparametric regression (SR) and simultaneous confidence bands may be used to examine differences between experimental groups in device monitoring studies (Maringwa et al, 2008), where group differences may not be constant over the monitoring time. Nested repeated measures (NRM) arise when subjects are monitored on multiple occasions in such studies; covariance structures for both the inner (monitoring time) and outer (occasion) RM factors should be included. NRM covariance structures have been proposed when using a parametric mean structure (Galecki, 1994). Although SR may provide a more flexible mean structure via penalized spline representation in the random effects, correlation between sets of spline coefficients has to be considered in this NRM setting. We propose a series of SR models that incorporate both levels of intrasubject correlation while preserving the ability to model the mean with a smooth nonparametric function. We examine the models through simulation studies. We illustrate the method for data collected from a sleep medicine study with parallel groups and NRM; 24-hour ambulatory blood pressure monitoring was taken before and after surgery for each subject.

email: rhonda.vandyke@cchmc.org

ANALYSIS OF LONGITUDINAL DATA USING ARMA(1,1) CORRELATION MODEL

Sirisha L. Mushti, Old Dominion University*
N. Rao Chaganty, Old Dominion University

Longitudinal or repeated measure data are an increasingly common feature of biomedical or clinical trials. Parsimonious one-parameter correlation models are often used for simplicity and ease of estimation, such as first-order auto-regressive (AR 1)), moving average (MA) or compound symmetric (CS) structures. In this research we consider the first-order autoregressive-moving average structure (ARMA(1,1)), which consists of two parameters and reduces to AR(1), MA and CS structures in special cases. We study positive definite ranges for the ARMA(1,1) model. Difficulties with maximum likelihood estimation due to the ARMA(1,1) structure are discussed, and two alternative methods for estimating the correlation parameters are presented, using pairwise likelihoods, and using composite bivariate likelihoods. Estimates obtained with these alternative methods are highly efficient compared to the maximum likelihood estimates in both asymptotic and small-sample cases, as shown through simulations. We illustrate with real-life data the use of this general ARMA(1,1) dependence structure.

email: smushti@odu.edu

GENERALIZED p-VALUE METHOD FOR TESTING ZERO VARIANCE IN LINEAR MIXED-EFFECTS MODELS

Haiyan Su, Montclair State University*
Xinmin Li, Shan Dong University of Technology
Hua Liang, University of Rochester
Wulin 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. We illustrate the proposed approach using an application from a nutritional study.

email: suh@mail.montclair.edu

VARIABLE SELECTION AND ESTIMATION FOR MULTIVARIATE PANEL COUNT DATA VIA THE SEAMLESS-L0 PENALTY

Haixiang Zhang, University of Missouri and University of Jilin, China*
Jianguo Sun, University of Missouri

Variable selection is fundamental to high-dimensional statistical analysis. In this article, we adopt the seamless-L0 penalty approach for variable selection with respect to multivariate panel count data. The proposed methodology selects variables and estimates regression coefficients simultaneously. Under certain regularity conditions, we show the consistency and asymptotic normality of the proposed estimator. Furthermore, the proposed method can be easily carried out with the Newton-Raphson algorithm. The performances of the procedure are evaluated by means of Monte Carlo simulation, and a data set from a motivating study of patients with skin cancers is analyzed as an illustrative example.

email: zhanghx09@mails.jlu.edu.cn

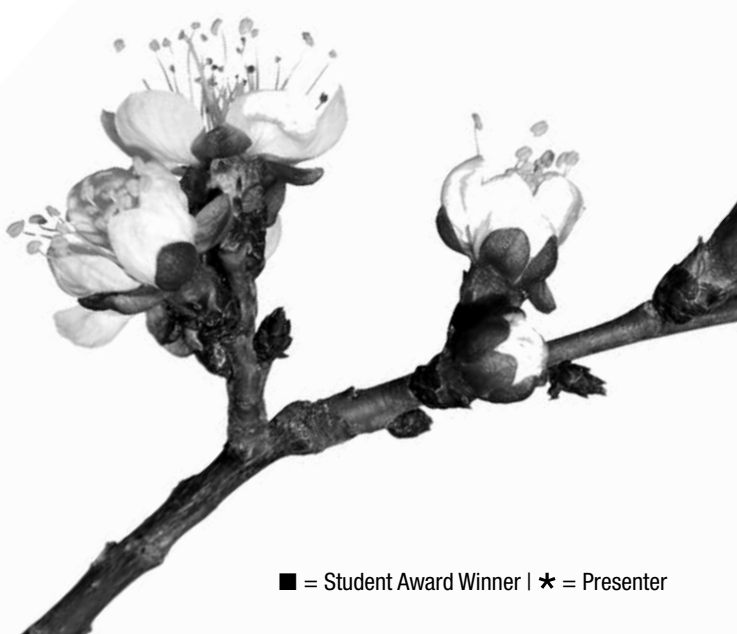
22. CORRELATED HIGH-DIMENSIONAL DATA

POSITIVE DEFINITE SPARSE ESTIMATORS OF HIGH-DIMENSIONAL COVARIANCE MATRICES

Adam J. Rothman, University of Minnesota*

Using convex optimization, we construct a sparse estimator of the covariance matrix that is positive definite and performs well in high-dimensional settings. A lasso-type penalty is used to encourage sparsity and a logarithmic barrier function is used to enforce positive definiteness. Consistency and convergence rate bounds are established as both the number of variables and sample size diverge. An efficient computational algorithm is developed and the merits of the approach are illustrated with simulations and a gene microarray data example.

email: arothman@umn.edu



■ = Student Award Winner | ✱ = Presenter

STATISTICAL MODELS FOR ANALYSIS OF HUMAN MICROBIOME DATA

Hongzhe Li, University of Pennsylvania*

Recent studies have suggested that the human gut microbiome performs numerous important biological functions and disorders of the microbiome are associated with many and diverse human disease processes. Systems biology approaches based on next generation sequencing technologies are now able to describe the gut microbiome at a detailed genetic and functional level,providing new insights into the importance of the gut microbiome in human health. In this talk, I will present several interesting statistical problems related to modeling the microbiome data, including methods for high dimensional regression for counts and compositional data and methods for investigating the dependency structure of the compositional data. I will illustrate these models and methods using a human gut microbiome study to linking diet to microbiome compositions.

email: hongzhe@upenn.edu

JOINT STATISTICAL MODELING OF MULTIPLE HIGH-DIMENSIONAL DATA

Yufeng Liu, University of North Carolina at Chapel Hill*

With the abundance of high dimensional data, shrinkage techniques are very popular for simultaneous variable selection and estimation. In this talk, I will present some new shrinkage techniques for joint analysis of multiple high dimensional data. Applications on cancer gene expression data and micro-RNA data will be presented.

email: yfliu@email.unc.edu

ON MAXIMUM LIKELIHOOD ESTIMATION OF MULTIPLE PRECISION MATRICES

Xiaotong Shen, University of Minnesota*
Yunzhang Zhu, University of Minnesota
Wei Pan, University of Minnesota

Consider the problem of estimation of multiple precision matrices in Gaussian graphical models, where the dependency structure is estimated between as well as within covariance matrices. Of particular interest is the grouping structure over these matrices as well as sparseness over and within matrices, which is characterized in terms of estimation of homogenous subgroups of elements and zero-elements of the matrices. This is well motivated by estimation of a change over a dynamic network in gene network analysis. We will discuss computational methods, in addition to some theory.

email: shenx002@umn.edu

23. CURRENT DEVELOPMENTS IN BAYESIAN CLINICAL TRIALS

BAYESIAN APPLICATIONS IN DRUG SAFETY EVALUATION

Amy Xia, Amgen, Inc.*

Safety assessment is critical in drug development and has called increasing attention recently. There are a few challenges in drug safety evaluation: How to detect unexpected adverse drug reactions while handling the multiplicity issue properly? How to synthesize data from different sources? How to deal with rare events? How to evaluate multidimensional, complex safety information as a whole? Bayesian hierarchical modeling offers many advantages and can be used to address these challenges. Specific applications include clinical trial signal detection, meta-experimental design and meta-analysis for rare adverse event data, joint modeling of longitudinal and time-to-event data, and continuously monitoring an adverse event of interest in which useful prior information can be incorporated into the decision making process. Practical considerations in dealing with data on safety are discussed. Advantages and challenges of Bayesian methods in these applications will be highlighted.

email: hxia@amgen.com

COMMENSURATE PRIORS FOR INCORPORATING HISTORICAL INFORMATION IN CLINICAL TRIALS USING GENERAL AND GENERALIZED LINEAR MODELS

Brian P. Hobbs, University of Texas MD Anderson Cancer Center
Daniel J. Sargent, Mayo Clinic
Bradley P. Carlin, University of Minnesota*

Assessing between-study variability in the context of conventional random-effects meta-analysis is notoriously difficult when incorporating data from only a small number of historical studies. In order to borrow strength, historical and current data are often assumed to be fully homogeneous a priori, with potentially drastic consequences for power and Type I error if the current data are revealed to conflict with the historical information. In this paper, we develop parametric empirical Bayes (EB) analogs of commensurate prior models (Hobbs et al., 2011 Biometrics) and evaluate their frequentist and Bayesian properties for incorporating patient-level historical data using general and generalized linear mixed models. The EB procedures, which estimate commensurability between the historical and concurrent control data, facilitate borrowing of strength from the historical data with only modest sacrifices in bias as compared to analyses that assume full homogeneity of the data sources. We illustrate with an example in a colon cancer trial setting where our proposed design produces more precise estimates of the model parameters. We also mention the approach's usefulness in adaptive randomization settings, where commensurability of historical controls might justify randomizing fewer current subjects to control.

email: brad@biostat.umn.edu

IDENTIFYING POTENTIAL ADVERSE EVENTS DOSE-RESPONSE RELATIONSHIPS VIA BAYESIAN INDIRECT AND MIXED TREATMENT COMPARISON MODELS

Haoda Fu, Eli Lilly & Company*
Karen L. Price, Eli Lilly & Company
Mary E. Nilsson, Eli Lilly & Company
Stephen J. Ruberg, Eli Lilly & Company

To help ensure patient safety for medical products, it is important to assess whether a potential adverse event dose response relationship exists, through combination of all the evidence from multiple clinical trials. The studies that need to be combined often include differing dose levels. In this paper, we propose three Bayesian indirect/mixed treatment comparison models to assess adverse event dose relationship. These three models are designed to handle binary responses and time to event responses. We apply the methods to real data sets and demonstrate that our proposed methods are useful in discovering relationships.

email: fuhaoda@gmail.com

24. CAUSAL INFERENCE AND MEASUREMENT ERROR

ANALYTIC RESULTS ON THE BIAS DUE TO NONDIFFERENTIAL MISCLASSIFICATION OF A CONFOUNDER

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

Suppose we are interested in the effect of a binary treatment on an outcome where that relationship is confounded by an ordinal confounder. We assume that the true confounder is not observed, rather we observe a nondifferentially mismeasured version of the confounder. We show that under certain monotonicity assumptions about the effect of the confounder on the treatment and on the outcome, the odds ratio, risk ratio, and risk difference calculated by standardizing by the mismeasured confounder will fall between the crude and the true effect measures on the corresponding scale. We present similar results for dichotomized confounders and for mismeasured continuous confounders. Under further assumptions, similar results also hold for multiple confounders.

email: ogburn@post.harvard.edu

MEASUREMENT BIAS IN CAUSAL INFERENCE: A GRAPH-BASED PERSPECTIVE

Judea Pearl, University of California at Los Angeles*

This paper addresses the problem of measurement errors in causal inference and highlights several algebraic and graphical methods for eliminating systematic bias due to such errors. In particular, the paper discusses the control of partially observable confounders in parametric and non parametric models and the computational problems of obtaining bias-free effect estimates in such models.

email: judea@cs.ucla.edu

MEDIATION ANALYSIS WHEN MEDIATOR IS MIS-MEASURED OR MIS-CLASSIFIED AND OUTCOME IS CONTINUOUS

Linda Valeri, Harvard University*
Tyler J. VanderWeele, Harvard University

Mediation analysis is a popular approach to studies in several fields to examine the extent to which the effect of an exposure to an outcome is through an intermediate and the extent to which is direct. When the mediator is mis-measured or misclassified the validity of mediation analysis can be severely undermined. The contribution of the present work is to study the effects of classical, non differential measurement error on the mediator in the estimation of direct and indirect causal effects when the outcome is continuous and exposure-mediator interaction can be present and to allow for the correction of such error. A correction along the lines of regression calibration with sensitivity analysis is proposed for which no validation samples or gold standard for the misclassified or mis-measured mediator are required. A strategy to effectively implement sensitivity analyses is proposed.

email: lvaleri@hsph.harvard.edu

AVERAGE CAUSAL EFFECT ESTIMATION ALLOWING COVARIATE MEASUREMENT ERROR

Yi Huang, University of Maryland, Baltimore County*
Xiaoyu Dong, University of Maryland, Baltimore County
Karen Bandeen-Roche, Johns Hopkins University
Cunlin Wang, U.S. Food and Drug Administration

The covariates are often measured with error in biomedical and policy studies, which is a violation of the strong ignorability assumption. The naive approach is to ignore the error and use the observed covariates in current propensity score framework for average causal effect (ACE) estimation. However, after extending the existing causal framework incorporating assumptions allowing errors-in-covariates, we showed that the naive approach typically produces biased ACE inference. In this talk, we developed a finite mixture model framework for ACE estimation with continuous

outcomes, which captures the uncertainty in propensity score subclassification from unobserved measurement error using the joint likelihood. The proposed approach will estimate the propensity score subgroup membership and subgroup-specific treatment effect jointly. Simulations studies and one real application (using the recent data from Infant Feeding Practice Study II) are used to show its performance and implementation. In summary, the proposed method extended the current propensity score subclassification approach to accommodate the cases where covariates are measured with errors.

email: yihuang@umbc.edu

25. TWO-PHASE ESTIMATION

INVESTIGATING ALTERNATIVE WAYS OF ESTIMATING THE PROPORTION OF A POPULATION WITH SERIOUS MENTAL ILLNESS FROM A TWO-PHASE SAMPLE

Phillip S. Kott, RTI International*
Dan Liao, RTI International
Jeremy Aldworth, RTI International

The National Survey of Drug Use and Health (NSDUH) uses a two-phase process to estimate the proportion of a population with serious mental illness (SMI). The first phase is the NSDUH itself, a large self-administered national survey containing a series of questions on personal drug use and mental health. A randomly chosen subsample of the annual NSDUH, the Mental Health Surveillance Survey (MHSS) is drawn, and respondents are clinically evaluated for SMI. A prediction model is fitted in this MHSS subsample with the clinical evaluations treated as the "gold standard" and then applied to the entire NSDUH sample. Currently, an unadjusted (model-based) cut-point estimator is computed by assigning an SMI status to everyone in the NSDUH based on a fitted logistic model. We investigated several potential alternatives to the unadjusted cut-point estimator above based on the same logistic-model fit with 2008 through 2010 NSDUH/MHSS data. We measured the standard errors of the competing estimators, including the error from estimating logistic-model parameters, using linearization and Fay's version of balanced repeated replication (Fay's BRR).

email: pkott@rti.org

EFFICIENT DESIGN AND INFERENCE FOR GENE X ENVIRONMENT INTERACTION, USING SEQUENCING DATA

Kenneth Rice, University of Washington*
Thomas Lumley, University of Auckland

The study of Gene x Environment interactions in humans present major challenges for both design and inference. In particular, while recently-available sequencing methods provide extremely detailed and reliable genetic data, they are still very expensive. This motivates careful design of genetic studies, weighting the allocation of sequencing resources to those with extreme trait values, and extreme environmental exposures. We describe how, particularly in multi-trait analyses, applying two-phase survey methods can provide efficiency gains over naive approaches, both in design and analysis. A similar transfer of sampling technology will illustrate how sampling based on known genotype can also provide efficiency gains, in certain circumstances.

email: kenrice@u.washington.edu

A MODEL ASSISTED APPROACH TO COMBINING DATA FROM TWO INDEPENDENT SURVEYS

Jae-kwang Kim, Iowa State University*
J.N.K Rao, Carleton University, Canada

Combining information from two or more independent surveys is a problem frequently encountered in survey sampling. We consider the case of two independent surveys, where a large sample from survey 1 collects only auxiliary information and a much smaller sample from survey 2 provides information on both the variables of interest and the auxiliary variables. We propose a model-assisted projection method of estimation based on a working model, but the reference distribution is design-based. We generate synthetic or proxy values of a variable of interest by first fitting the working model, relating the variable of interest to the auxiliary variables, to the data from survey 2 and then predict the variable of interest associated with the auxiliary variables observed in survey 1. The projection estimator of a total is simply obtained from the survey 1 weights and associated synthetic values. We identify the conditions for the projection estimator to be asymptotically unbiased. Domain estimation using the projection method is also considered. Replication variance estimators are obtained by augmenting the synthetic data file for survey 1 with additional synthetic columns associated with the columns of replicate weights. Results from a simulation study are presented.

email: jkim@iastate.edu

26. SEMI-COMPETING RISKS

BAYESIAN GAMMA FRAILTY MODELS FOR SURVIVAL DATA WITH SEMI-COMPETING RISKS AND TREATMENT SWITCHING

Yuanye Zhang, University of Connecticut
Ming-Hui Chen, University of Connecticut*
Joseph G. Ibrahim, University of North Carolina at Chapel Hill
Donglin Zeng, University of North Carolina at Chapel Hill
Qingxia Chen, Vanderbilt University
Zhiying Pan, Amgen Inc.
Xiaodong Xue, Amgen Inc.

In this paper, we propose a class of semi-competing risk gamma-frailty survival models to account for treatment switching and dependence between disease progression time and survival time. Properties of the proposed model are examined and an efficient Gibbs sampling algorithm is developed. Deviance Information Criterion (DIC) with an appropriate deviance function and Logarithm of the Pseudomarginal Likelihood (LPML) are constructed in order to compare the proposed model to the semi-competing risk transition model. An extensive simulation study is carried out to examine the performance of DIC and LPML and as well as the frequentist performance of posterior estimates. The proposed method is also applied to analyze data from the panitumumab study.

email: ming-hui.chen@uconn.edu

QUANTILE REGRESSION METHODS FOR SEMI-COMPETING RISKS DATA

Limin Peng, Emory University*

Semi-competing risks data are frequently encountered in clinical studies that involve multiple terminating and nonterminating events. Handling such data is often complicated by the non-independent relationship between the endpoint of interest and its competing events. According to data scenarios, one may opt to analyses oriented to crude quantities or net quantities in order to generate meaningful scientific implications. In this talk, I will present an overall framework for conducting quantile regression methods in the presence of semi-competing risks. As an alternative regression strategy to traditional regression methods in survival analysis, quantile regression can produce a more comprehensive picture for the association between event time outcomes and covariates. I will present sensible modeling and inferential procedures tailored to semi-competing settings. Data examples will be presented to illustrate the utility of the developed methods.

email: lpeng@sph.emory.edu

NONPARAMETRIC CAUSE-SPECIFIC ASSOCIATION ANALYSES OF MULTIVARIATE UNTIED OR TIED COMPETING RISKS DATA

Hao Wang, University of Pittsburgh
Yu Cheng, University of Pittsburgh*

We extend the bivariate hazard ratio (Cheng and Fine, 2008) to multivariate competing risks data and show that it is equivalent to the cause-specific cross hazard ratio in Cheng et al. (2010). Two nonparametric approaches are proposed to estimate the two equivalent association measures. One extends the plug-in estimator in Cheng and Fine (2008) and the other adapts the pseudo likelihood estimator for bivariate survival data (Clayton, 1978) to multivariate competing risks data. Their asymptotic properties are established by using empirical processes techniques. We compare the extended plug-in and pseudo likelihood estimators with the existing U statistic by simulations and show that the three methods have comparable performance when the data have no tied events. However, all the three estimators are biased in the presence of rounding errors. We hence propose a modified U statistic to take into account tied observations, which clearly outperforms the other estimators when there are rounding errors. All methods are applied to the Cache County Study to examine familial associations in dementia among this aging population. We recommend using the simple plug-in estimator for untied data, and using the modified U statistic for tied data.

email: yucheng@pitt.edu

ESTIMATION OF TIME-DEPENDENT ASSOCIATION FOR BIVARIATE FAILURE TIMES IN THE PRESENCE OF A COMPETING RISK

Jing Ning, University of Texas MD Anderson Cancer Center*
Karen Bandeen-Roche, Johns Hopkins University

This talk targets the estimation of a time-dependent association measure for bivariate failure times, the conditional cause-specific hazard ratio, which is a generalization of the conditional hazard ratio to accommodate competing risks data. We model the conditional cause-specific hazard ratio as a parametric regression function of time, event causes and other covariates, and leave all other aspects of the joint distribution of the failure times unspecified. We develop a pseudo-likelihood estimation procedure for model fitting and inference and establish the asymptotic properties of the estimators. We assess the finite-sample properties of the proposed estimators against the estimators obtained from a moment-based estimating equation approach. Data from the Cache County study on dementia are used to illustrate the proposed methodology.

email: jning@mdanderson.org

27. GRADUATE STUDENT AND RECENT GRADUATE COUNCIL INVITED SESSION: CAREERS IN BIOSTATISTICS

THE GRADUATE STUDENT AND RECENT GRADUATE COUNCIL

Hormuzd Katki, National Cancer Institute, National Institutes of Health

The ENAR Regional Advisory Board (RAB) proposes to establish a Graduate Student and Recent Graduate Council (GSRGC) to allow ENAR to better serve the special needs of students and recent graduates. I will describe example activities we envision the GSRGC participating in, how it would be constituted, and how it will interface with RAB and ENAR leadership. We are looking for students and recent graduates who would like to serve on the GSRGC.

email: katkih@mail.nih.gov

ARE YOU A HEDGEHOG OR A FOX? BRIEF COMMENTS ON A CAREER IN STATISTICAL CONSULTING

Jennifer Schumi, Statistics Collaborative

In graduate school, we develop our technical skills in statistics through course work and independent research. Some statisticians continue that focused methodological work after completing their degrees, while others pursue a more applied path collaborating with physicians, scientists, regulators, and policy makers, among others. While our technical skills get us in the door, success as applied statisticians requires us to draw upon many other talents. Through examples of projects in public health and clinical trials, I will discuss some of the facets of my career as a practicing consulting statistician.

email: jennifer@statcollab.com

CAREERS OF STATISTICIANS AND BIOSTATISTICIANS IN THE GOVERNMENT

Telba Irony, U. S. Food and Drug Administration

Statisticians are extremely respected professionals who play a crucial role in the Department of Health and Human Services and in particular, in the approval process of medical treatments and diagnostics by the Food and Drug Administration. In this presentation we will discuss the responsibilities of statisticians in the government, highlighting the need of communication and collaborative skills in addition to technical statistical skills. The statistician in government is a problem solver, who must be interested in science and teaching, and could aspire to leadership positions. The statistician can be a force that spurs innovation, not only in science and medicine, but also in statistical techniques and decision making processes. We will also discuss the career ladder of statisticians in the government and will be open to questions and comments from the audience.

email: Telba.irony@fda.hhs.gov

LIVING AND WORKING IN ACADEMIA POST GRADUATION

Kimberly L. Drews, The George Washington University

This presentation will provide information about careers in academia. It will explain the metric for success in an academic work environment and provide insight into the three components of the metric for this success: education, service, and scholarship. It will also cover how these three components can vary in importance based on academic position, department and university. A few keys for success will also be presented.

email: kdrews@bsc.gwu.edu

28. STATISTICAL CHALLENGES OF SPATIAL MULTI-POLLUTANT DATA IN ENVIRONMENTAL EPIDEMIOLOGY

METHODS FOR SPATIALLY-VARYING MEASUREMENT ERROR IN AIR POLLUTION EPIDEMIOLOGY

Stacey E. Alexeeff, Harvard School of Public Health*
Raymond J. Carroll, Texas A&M University
Brent A. Coull, Harvard School of Public Health

Land use regression models can improve exposure assessment for traffic-related air pollutants, especially compared to previous approaches using central monitoring sites. However, a health effect analysis that uses predicted values from an exposure model results in exposure misclassification because the predicted exposures are not the true exposures. Thus, the health model can be viewed as containing a covariate with measurement error, and the magnitude of the error may vary by location. Statistical methodology should properly account for the uncertainty associated with modeled predictions. We explore simulation based approaches, focusing on functional models which place minimal assumptions on the distribution of the exposures. These approaches provide a more flexible correction method that could be applied to many different exposure prediction models, rather than being a model-specific correction method. We apply the proposed methods to an analysis of air pollution and birthweight in Boston.

email: salexeeff@fas.harvard.edu

REDUCED BAYESIAN HIERARCHICAL MODELS: ESTIMATING HEALTH EFFECTS OF SIMULTANEOUS EXPOSURE TO MULTIPLE POLLUTANTS

Jennifer F. Bobb, Johns Hopkins Bloomberg School of Public Health*
Francesca Dominici, Harvard School of Public Health
Roger D. Peng, Johns Hopkins Bloomberg School of Public Health

Quantifying the health effects associated with simultaneous exposure to many air pollutants is now a research priority of the US EPA. Bayesian hierarchical models (BHM) have been extensively used in multisite time series studies of air pollution and health to estimate health effects of a single pollutant adjusted for potential confounding of other pollutants and other time-varying factors. However, when the scientific goal is to estimate the health effects of many pollutants jointly, a straightforward application of BHM is challenged by the need to specify a random-effect distribution on a high-dimensional vector of nuisance parameters, which often do not have an easy interpretation. In this paper we introduce an improved BHM formulation, which we call 'reduced BHM,' aimed at analyzing clustered data sets in the presence of a large number of random effects that are not of primary scientific interest. In simulation studies we show that the reduced BHM performs comparably to the full BHM in many scenarios, and even performs better in some cases. Methods are applied to estimate location-specific and overall relative risks of cardiovascular hospital admissions associated with simultaneous exposure to elevated levels of particulate matter and ozone in 51 US counties during the period 1999-2005.

email: jenniferfederbobb@gmail.com

SPATIAL VARIABLE SELECTION METHODS FOR ESTIMATING HEALTH EFFECTS OF SPECIATED PARTICULATE MATTER

Laura F. Boehm, North Carolina State University*
Francesca Dominici, Harvard School of Public Health
Brian J. Reich, North Carolina State University
Montserrat Fuentes, North Carolina State University

Previous research has suggested a connection between ambient particulate matter (PM) exposure and acute health effects, but the effect size varies across the United States. Variability in the effect may partially be due to differing community level exposure and health characteristics, but also due to the chemical composition of PM which is known to vary greatly by location and over time. The scientific goal is to identify particularly toxic chemical components of this chemical mixture. Because of the large number of potentially highly correlated components, we must incorporate some regularization into a statistical model. We assume that at each location, regression coefficients come from a mixture model, with the flavor of stochastic search variable selection, but utilize a copula to share information about variable inclusion and effect magnitude across locations. The model will differ from current spatial variable selection techniques by simultaneously describing local and global variable selection. The model will be applied to fine PM (PM <2.5 μm), measured at 118 counties nationally, and cardiovascular emergency room admissions among Medicare patients, over the period 2000-2008.

email: lfboehm@ncsu.edu

BAYESIAN SPATIALLY-VARYING COEFFICIENT MODELS FOR ESTIMATING THE TOXICITY OF THE CHEMICAL COMPONENTS OF FINE PARTICULATE MATTER

Yeonseung Chung, Korea Advanced Institute of Science and Technology*
Francesca Dominici, Harvard School of Public Health
Michelle Bell, Harvard School of Public Health
Brent Coull, Harvard School of Public Health

Several studies have reported associations between long-term exposure to ambient fine particulate matter (PM2.5) and mortality. However, it is not much explored which chemical constituents determine the toxicity of PM2.5. The health effects of long-term exposure to PM2.5 vary across different locations and such spatial heterogeneity in health responses to long-term PM2.5 may be explained by the chemical composition of PM2.5. In this research, we propose a statistical model to investigate the spatially-varying (SV) health effects of long-term PM2.5 and the effect modification by the chemical components simultaneously. We use a Bayesian SV coefficient Poisson regression to quantify the spatially-heterogeneous toxicity of long-term PM2.5 on mortality; (2) we regress the chemical component levels on the SV coefficients to identify the components that modify the PM2.5 toxicity. Applying the proposed model to the US Medicare Cohort Air Pollution Study (MCAPS) data, we encounter a missing value problem because the chemical component data is sparser than the PM2.5 data. We adopt a Gaussian spatial process as a prediction model for the missing component levels. Using the complete case data, we conduct cross-validation studies to examine the prediction performance and the impact of prediction on the health effects estimates.

email: dolyura@kaist.edu

A BIVARIATE SPACE-TIME DOWNSCALER UNDER SPACE AND TIME MISALIGNMENT

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

Ozone and particulate matter are co-pollutants that have long been associated with increased public health risks. Information on concentration levels for both pollutants come from two sources: monitoring sites and output from complex numerical models that produce concentration surfaces over large spatial regions. We offer a fully-model-based approach for fusing these two sources of information for the pair of co-pollutants which is computationally feasible over large spatial regions and long periods of time. Due to the association between concentration levels of the two environmental contaminants, it is expected that information regarding one will help improve prediction of the other. Misalignment is an obvious issue since the monitoring networks for the two contaminants only partly intersect and because the collection rate for particulate matter is typically less frequent than ozone.

email: berrocal@umich.edu

29. SAMPLE SIZE ADJUSTMENTS FOR CLINICAL TRIALS WITH MULTIPLE COMPARISONS

SAMPLE SIZES FOR TRIALS INVOLVING MULTIPLE CORRELATED MUST-WIN COMPARISONS

Steven A. Julious, University of Sheffield*
Nikki E. McIntyre, AstraZeneca

In Clinical trials involving multiple comparisons of interest, the importance of controlling the trial Type I error is well-understood and well-documented. Moreover, when these comparisons are themselves correlated, methodologies exist for accounting for the correlation in the trial design, when calculating the trial significance levels. Less well-documented is the fact that there are some circumstances where multiple comparisons affect the Type II error rather than the Type I error, and failure to account for this, can result in a reduction in the overall trial power. In this talk we describe sample size calculations for clinical trials involving multiple correlated comparisons, where all the comparisons must be statistically significant for the trial to provide evidence of effect, and show how such calculations have to account for multiplicity in the Type II error. We begin with a simple case of two comparisons assuming a bivariate Normal distribution, show how to factor in correlation between comparisons and then generalise our findings to situations with 2 or more comparisons. These methods are easy to apply, and we demonstrate how accounting for the multiplicity in the Type II error leads, at most, to modest increases in the sample size.

email: S.A.Julious@Sheffield.ac.uk

SAMPLE SIZES ACCOUNTING FOR MULTIPLICITY: IMPORTANCE IN PHASE 2

Brian L. Wiens, Alcon Laboratories, Inc.*
Srichand Jasti, Alcon Laboratories, Inc.
John W. Seaman, Alcon Laboratories, Inc.

We consider design considerations for a phase 2 study in which the endpoint that will support registration is not determined before the phase 2 study begins. Multiple endpoints, any of which could support registration in a phase 3 study, are assessed in the phase 2 study. Assessment of multiple endpoints in the phase 2 study requires control of the type I error rate. Further, estimation of treatment effect is subject to bias if the endpoint with larger treatment effect is chosen for the phase 3 study, resulting in overestimation of power. We consider analysis methods that account for multiple endpoints to, first, demonstrate an effect and, second, to choose a primary endpoint for the phase 3 trial.

email: brian.wiens@alconlabs.com

POWER AND SAMPLE SIZE DETERMINATION IN CLINICAL TRIALS WITH TWO-CORRELATED RELATIVE RISKS

Toshimitsu Hamasaki, Osaka University Graduate School of Medicine*
Scott Evans, Harvard University School of Public Health
Tomoyuki Sugimoto, Hirosaki University
Takashi Sozu, Kyoto University School of Public Health

The effects of interventions are multi-dimensional (e.g., benefits and harms). Co-primary endpoints offer an attractive design feature in clinical trials as they capture a more complete characterization of the effect of an intervention. For example in cancer trials, overall survival is often of primary interest, but relapse-free or progression-free survival is also important. Trials of co-morbidities may also utilize co-primary endpoints, e.g., a trial evaluating therapies to treat Kaposi's sarcoma (KS) in HIV-infected individuals may have: (1) the time to KS progression and (2) the time to HIV virologic failure, as co-primary endpoints. In the presentation, we discuss power and sample size determination for comparative clinical trials with two correlated relative risks to be evaluated as primary contrasts. We consider two situations: (a) where the objective is to provide statistical significance in favor of the test treatment compared with the control treatment for all of the relative risks, and (b) where the objective is to demonstrate statistical significance for at least one relative risk. We evaluate how the required sample size and power vary as a function of the correlation between the outcomes.

email: hamasakt@medstat.med.osaka-u.ac.jp

TEST AND POWER CONSIDERATIONS FOR MULTIPLE ENDPOINT ANALYSES USING SEQUENTIALLY REJECTIVE GRAPHICAL PROCEDURES

Frank Bretz, Novartis*
Willi Maurer, Novartis
Ekkehard Glimm, Novartis

A variety of powerful test procedures are available for the analysis of clinical trials addressing multiple objectives, such as comparing several treatments with a control, assessing the benefit of a new drug for more than one endpoint, etc. Graphical approaches have recently been proposed that facilitate the derivation and communication of tailored multiple test strategies. In this presentation we discuss suitable power measures for clinical trials with multiple primary and/or secondary objectives. We discuss the importance of choosing suitable weights for the underlying closed test procedures and how to optimally propagate local significance levels to meet the study objectives. We use a generic example to illustrate the results.

email: frank.bretz@novartis.com

SAMPLE SIZE OF THOROUGH QTc CLINICAL TRIAL ADJUSTED FOR MULTIPLE COMPARISONS

Yi Tsong, U.S. Food and Drug Administration*
Xiaoyu Dong, University of Maryland at Baltimore County

A thorough QT trial is typically designed to test for two set of hypotheses. The primary set of hypotheses is for demonstrating that the test treatment will not prolong QT interval. The second set of hypotheses is to demonstrate the assay sensitivity of the positive control treatment in the study population. Both analyses require multiple comparisons by testing the treatment difference measured repeatedly at multiple selected time points. Tsong et al (2010a) indicated that for the prolongation testing, it involves union-intersection test that lead to the reduction of study power. Tsong et al (2010b) indicated also that the assay sensitivity analysis is carried out using intersection-union test that lead to the inflation of the family-wise type I error rate and requires type I error rate adjustment to control the family-wise type I error rate. Zhang and Machado (2008) proposed the sample size calculation of test-placebo QT response difference based on simulation with a multivariate normal distribution model. Theses simulation results are limited to potential general generalization to various advanced and adaptive designs of TQT trials (Tsong (2012)). We propose a sample size determination using power equation based on multivariate normal distribution but adjusted for multiple comparisons.

email: yi.tsong@fda.hhs.gov

30. ADAPTIVE DESIGN/ADAPTIVE RANDOMIZATION

PLATFORM-BASED CLINICAL TRIAL DESIGNS FOR EFFICIENT DRUG DEVELOPMENT STRATEGIES

Brian P. Hobbs, University of Texas MD Anderson Cancer Center*
J. Jack Lee, University of Texas MD Anderson Cancer Center

Conventionally, evaluation of novel therapeutic agents in phase II involves separate single-arm studies designed to assess a small number of pre-specified agents sequentially. In this paper we propose "platform-based" clinical trial designs that facilitate simultaneous assessments of multiple agents for screening drugs' efficacy efficiently, and treat more patients with more effective treatments during the trial. The platform approach establishes a program involving a single protocol that facilitates seamless modifications to the study arms as poorly and well performing agents are dropped or graduated, respectively, and replaced by new agents using Bayesian group sequential methods. Simulation is used to evaluate the operating characteristics of platform designs that simultaneously compare up to seven treatment arms to standard therapy over a period of five years. In addition, we compare designs that allocate patients to study arms using equal randomization, Bayesian adaptive randomization, as well as two novel adaptive randomization methods: precision adaptive randomization, and weighted Bayesian adaptive randomization. On average the platform designs are shown to screen twice as many agents and provide 52% better overall response when compared to conventional designs that screen agents at a time.

email: bphobbs@mdanderson.org

A BAYESIAN DECISION-THEORETIC SEQUENTIAL-RESPONSE ADAPTIVE RANDOMIZATION DESIGN

Fei Jiang, Rice University*
J. Jack Lee, University of Texas MD Anderson Cancer Center
Peter Mueller, University of Texas at Austin

We propose a class of phase II clinical trial designs with sequential stopping and adaptive treatment allocation to evaluate treatment efficacy. Our discussions are based on two-arm (control and experimental treatment) designs with binary end points. The designs combine the Bayesian decision-theoretic sequential approach with the adaptive randomization procedures in order to achieve the efficient and ethical goals simultaneously. The design parameters represent the costs of different decisions, e.g. the decisions for stopping or continuing the trials. The parameters enable us to incorporate the actual costs of the decisions in practice. The proposed designs allow the clinical trials to stop early for either efficacy or futility. Furthermore, the designs assign more patients to better treatment arms by applying the adaptive randomization procedures. We develop an algorithm based on the constrained backward induction and forward

simulation to implement the designs. The algorithm overcomes the computational issues of the standard methods, thereby making our approach applicable. The designs result in the trials with desirable operating characteristics under the simulated settings. Moreover, the designs are robust with respect to the response rate of the control group.

e-mail: homebovine@hotmail.com

EXTENDING THE TITE CRM TO MULTIPLE OUTCOMES

Joseph S. Koopmeiners, University of Minnesota*

In traditional phase 1 oncology trials, the safety of a new chemotherapeutic agent is tested in a dose escalation study to identify the maximum tolerated dose, which is defined as the highest dose with acceptable toxicity. An alternate approach is to jointly model toxicity and efficacy and allow dose finding to be directed by a pre-specified tradeoff between efficacy and toxicity. With this goal in mind, several phase 1 designs have been proposed to jointly model toxicity and efficacy in phase 1 dose escalation studies. A factor limiting the use of these designs is that toxicity and efficacy must be observed in a timely manner. This is particularly problematic for the efficacy outcome, which is often measured over a longer timeframe than the toxicity outcome. One approach to overcoming this problem is to model toxicity and efficacy as time-to-event outcomes. We propose a phase 1 dose escalation study that jointly models toxicity and efficacy as time-to-event outcomes by extending the time-to-event CRM to accommodate multiple outcomes. The operating characteristics of our proposed design are evaluated by simulation and compared to the operating characteristics for existing phase 1 designs that consider toxicity and efficacy as binary outcomes.

e-mail: koopm007@umn.edu

A BAYESIAN ADAPTIVE ALLOCATION METHOD FOR CLINICAL TRIALS WITH DUAL OBJECTIVES

Roy T. Sabo, Virginia Commonwealth University*
Ghalib Bello, Virginia Commonwealth University
Lauren Grant, Virginia Commonwealth University
Cathy Roberts, Virginia Commonwealth University
Amir A. Toor, Virginia Commonwealth University
John M. McCarty, Virginia Commonwealth University

This research focuses on producing adaptive allocation proportions for Phase II clinical trials with two primary response endpoints. Special attention is given to utilizing Bayes methods to gauge treatment performance (i) as compared to hypothesized standards, (ii) based on inter-treatment comparisons, and (iii) a hybrid of those two cases. Simulation studies were conducted to show the behavior of the adaptive allocation weights under various clinical settings, as well as to show the effects of adaptive randomization on early study termination. The choice of using either posterior or predictive probabilities to calculate the allocation weights is also studied. Retrospective data consisting of 373 autologous transplant patients undergoing stem cell mobilization from the VCU Massey Cancer Center are used to simulate a prospective clinical trial featuring adaptive randomization for dual efficacy and futility outcomes; in this example, a definitive conclusion is reached well before the planned end of the study. This Bayesian adaptive approach to dual-outcome clinical trial design has the potential to randomize patients into more efficacious, less toxic and less futile treatment regimens, can reduce the time needed to reach a study's conclusion, and can ultimately reduce the amount of resources needed to conduct a trial.

e-mail: rsabo@vcu.edu

A SIMULATION STUDY TO DECIDE THE TIMING OF AN INTERIM ANALYSIS IN A BAYESIAN ADAPTIVE DOSE-FINDING STUDIES WITH DELAYED RESPONSES

Xiaobi Huang, Merck & Co., Inc.*
Haoda Fu, Eli Lilly and Company

The use of Bayesian adaptive design draws considerable attentions in dose-finding studies to improve trial efficiency. In certain therapeutic areas as diabetes and obesity, studies take weeks or months for a drug effect to emerge. Thus at the time of an interim analysis, the majority of patients have not completed the study. Fu and Manner (2010) proposed a Bayesian prediction model to incorporate these patients to the interim analysis. It is crucial to understand when to conduct the interim analysis, since an early adaptation may suffer from the lack of information while a later adaptation may cause loss of efficiency. In this paper, we conduct a simulation study to evaluate the timing of interim analysis under different dosing regimens and recruiting rates. These results provide general recommendations on how to decide the timing of an interim analysis.

e-mail: xiaobih@umich.edu

A TRIVARIATE CONTINUAL REASSESSMENT METHOD FOR PHASE I/ II TRIALS OF TOXICITY, EFFICACY, AND SURROGATE EFFICACY

Wei Zhong, University of Minnesota*
Joseph S. Koopmeiners, University of Minnesota
Bradley P. Carlin, University of Minnesota

Recently, many Bayesian methods have been developed for dose-finding when simultaneously modeling both toxicity and efficacy outcomes in a blended phase I/II fashion. A further challenge arises when all the true efficacy data cannot be obtained quickly after the treatment, so that surrogate markers are instead used (e.g, in cancer trials). We propose a framework to jointly model the probabilities of toxicity, efficacy and surrogate efficacy given a particular dose. Our trivariate binary model is specified as a

composition of two bivariate binary submodels. In particular, we extend the bCRM approach (Braun, 2002), as well as utilize the Gumbel copula of Thall and Cook (2004). Our proposed Bayesian trivariate dose-finding algorithm utilizes all the available data at any given time point, and can flexibly stop the trial successfully improve dosage targeting efficiency and guard against excess toxicity over a variety of true model settings and degrees of surrogacy. We conclude with a discussion of potential future work, especially regarding the use of more flexible link functions in our probability models.

e-mail: zhong038@umn.edu

31. BIOMARKERS I

LOGNORMAL AND GAMMA MODELS TO ESTIMATE MEANS FOR SKEWED BIOMARKER DATA SUBJECT TO ASSAY POOLING

Emily M. Mitchell, Emory University*
Robert H. Lyles, Emory University
Neil J. Perkins, National Institute of Child Health and Development, National Institutes of Health
Enrique F.Schisterman, National Institute of Child Health and 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 highlights the utility of maximum likelihood estimation under normality assumptions when pooling is conducted for a continuous outcome or predictor variable. Many public health studies, however, involve skewed data, often assumed to be log-normally distributed, which complicates the estimation procedure. Several methods have been proposed to approximate the distribution of a sum of lognormal variates, with an emphasis placed on moment-matching. In particular, this study focuses on a previously proposed approximation based on the modified power lognormal distribution. We use simulations to compare mean and variance estimation using this approximation technique with an exact convolution integral-based approach, as well as with estimation under an assumed gamma distribution. Pool sizes of 2 and 3 are considered, and data simulated from both lognormal and gamma distributions are analyzed to determine the effect of model misspecification on the corresponding estimates. Sensitivity of each strategy to changes in the overall sample size and the true mean and variance parameters is also investigated. We apply these methods to the analysis of cytokine data for which individual as well as pooled samples are available.

email: emitch8@emory.edu

PROSPECTIVE POOLING FOR DISCRETE SURVIVAL OUTCOME

Paramita Saha Chaudhuri, Duke University School of Medicine*
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

Pooled exposure analysis has become a very useful technique especially when the exposure assay is expensive or limited volume of specimen is available for assaying. Saha and Weinberg (2010) extended the pooled exposure analysis for a discrete-time survival outcome such as time-to-pregnancy. Two limitations exist. First the pooled exposure cannot be reused for another disease. Subjects concordant for cancer may not be concordant for heart disease, hence pooling needs to be done afresh with each new disease studied. Moreover, the analysis assumes a logistic model that may not hold in practice. We introduce a multiple imputation approach that uses a prospective pooling where subjects are grouped at the outset of the study without restricting the grouping within same outcome stratum addressing the first limitation. A flexible modeling approach is employed to account for a general risk model. We show that for a discrete-time survival outcome, this approach can be employed to test for the exposure effect on the time to outcome and can be used to estimate the exposure effect under certain conditions. We demonstrate this approach via extensive simulation studies and real data example.

email: paramita.sahachaudhuri@duke.edu

AN APPLICATION OF THE RARE AND WEAK MODEL IN BIOMARKER DISCOVERY IN PROTEOMICS STUDY

Xia Wang, University of Cincinnati*
Nell Sedransk, National Institute of Statistical Sciences

In clinical proteomics, biomarker discovery studies focuses on detecting differentially expressed proteins in cancer and in chronic diseases. Recently label-free methods have gained its popularity as a technique for relative quantitation in mass spectrometry-based proteomics. In label-free proteomics, the intensities of peptides are measured, from which the abundance of proteins are inferred. To model the correlated peptide intensities data, a functional mixed model is applied to the data ordered the retention times and the m/z of the precursor peptides. As a general situation, only a small proportion of the peptides are expected to differ between samples. The study employs the rare and weak model in feature selection, which is particular useful for data with high dimension and small sample size. This proposed approach is applied in CPTAC data with yeast samples spiked with known proteins as well as E.coli datasets (Finney et al. 2008).

email: xiawang.z@gmail.com

META-REGRESSION MODELS TO DETECT BIOMARKERS CONFOUNDED BY STUDY-LEVEL COVARIATES IN MAJOR DEPRESSIVE DISORDER MICROARRAY DATA

Xingbin Wang, University of Pittsburgh*
Etienne Sibille, University of Pittsburgh
George C. Tseng, University of Pittsburgh

Meta-analysis has become popular in the biomedical research because it generally can increase statistical power and provide validated conclusions. However, its result is often biased due to the heterogeneity. Meta-regression has been a useful tool for exploring the source of heterogeneity among studies in a meta-analysis. In this paper, we will explore the use of meta-regression in microarray meta-analysis. To account for heterogeneities introduced by study-specific features such as sex, brain region and array platform in the meta-analysis of major depressive disorder (MDD) microarray studies, we extended the random effects model (REM) for genomic meta-regression, combining eight MDD microarray studies. Due to the small number of studies, we proposed meta-regression with variable selection by Bayesian Information Criterion (BIC) such that at most one study-specific variable is included in the model. The result shows increased statistical power to detect gender-dependent and brain-region-dependent biomarkers that traditional meta-analysis methods cannot detect. The identified gender-dependent markers have provided new biological insights as to why females are more susceptible to MDD and the result may lead to novel therapeutic targets.

email: xingbinw@gmail.com

ESTIMATION OF C-INDEX FOR CENSORED BIOMARKER DATA IN COX PROPORTIONAL HAZARD MODEL

Yeonhee Kim, INC Research*
Lan Kong, Penn State Hershey College of Medicine

In recent few years, an increasing number of biomarkers has been investigated to predict survival time for patients with acute or chronic diseases. A widely used method for these types of research is Cox proportional hazard model and its discrimination power is often evaluated by Harrell's C-index. When some of the biomarker measurements are censored due to a detection limit of given assay, however, current estimation methods may yield erroneous results. We propose a likelihood-based approach to estimate the Cox proportional hazard model in the presence of censored biomarker covariate and derive the estimator of C-index that represents the potential discrimination power of a biomarker when there is no detection limit. Simulation study demonstrates that the proposed method outperforms over the simple substitution methods where the censored observations are replaced by an arbitrary constant. Our method is applied to a biomarker study to predict time to recovery from acute kidney injury.

email: yhkimbani@gmail.com

LOGISTIC REGRESSION ANALYSIS OF BIOMARKER DATA SUBJECT TO POOLING AND DICHOTOMIZATION

Zhiwei Zhang, U.S. Food and Drug Administration*
Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Robert H. Lyles, Emory University
Bhramar Mukherjee, University of Michigan

There is growing interest in pooling specimens across subjects in epidemiologic studies, especially those involving biomarkers. This paper is concerned with regression analysis of epidemiologic data where a binary exposure is subject to pooling and the pooled measurement is dichotomized to indicate either that no subjects in the pool are exposed or that some are exposed, without revealing further information about the exposed subjects in the latter case. The pooling process may be stratified on the disease status (a binary outcome) and possibly other variables but is otherwise assumed random. Methods are proposed for estimating parameters in a prospective logistic regression model, and illustrated with data from a population-based case-control study of colorectal cancer. Simulation results show that the proposed methods perform reasonably well in realistic settings and that pooling can lead to sizable gains in cost-efficiency. Recommendations are made with regard to the choice of design for pooled epidemiologic studies.

email: zhiwei.zhang@fda.hhs.gov

32. CAUSAL INFERENCE

CAUSAL INFERENCE WITH TREATMENT DELAY: EVALUATING MEDICATION USE IN WOMEN WITH HIGH RISK FOR PRETERM BIRTH VIA PROPENSITY SCORE MATCHING

Erinn Hade, The Ohio State University
Bo Lu, The Ohio State University*
Hong Zhu, The Ohio State University

In many observational clinical studies, the patients do not always get the treatment at the intended time. The intervention may be delayed for various reasons, including insurance coverage, scheduling issues or simply that the patients need more time to think about it. Weekly injections of 17 alpha-hydroxyprogesterone caproate (17P) is a standard treatment for women with high risk for preterm birth. Although positive results for clinical use

of 17P have been shown, the mechanisms by which it works have not been fully determined and it remains unclear whether weekly 17P injections prevent the cervix from shortening. In an observational cohort of women with preterm birth history, many did not initiate 17P immediately following their first visits (though offered), but received the treatment at later visits. We develop an innovative propensity score matching approach to assess both the instantaneous and long-term effect of 17P use by taking advantage of the treatment delay information.

email: blu@cph.osu.edu

A NEW DISTRIBUTION-FREE APPROACH FOR LONGITUDINAL MEDIATION ANALYSIS WITH NON-CONTINUOUS OUTCOMES AND MEDIATORS

Douglas D. Gunzler, Case Western Reserve University*

Mediation analysis constitutes an important part of treatment study to identify the mechanisms by which an intervention achieves its effect. Structural equation model (SEM) is a popular framework for modeling such a causal relationship. However, current methods impose various restrictions on the study designs and data distributions, greatly limiting the utility of the information they provide in real study applications. This problem is only magnified with non-continuous outcomes and mediators. In particular, for the binary outcome or mediator case, existing models are problematic even under complete data. We propose a new approach to address the limitations of current SEM within the context of longitudinal mediation analysis for binary outcomes and mediators by utilizing a class of functional response models (FRM). Being distribution-free, the FRM-based approach does not impose any parametric assumption on data distributions. In addition, by extending the inverse probability weighted (IPW) estimates to the current context, the FRM-based SEM provides valid inference for longitudinal mediation analysis under the two most popular missing data mechanisms; missing completely at random (MCAR) and missing at random (MAR). We illustrate the approach with both real and simulated data, resulting in a comprehensive logit link model for ease of interpretation and application.

email: dgunzler@metrohealth.org

LARGE SAMPLE PROPERTIES OF MULTIPLICATIVE TREATMENT EFFECT ESTIMATE USING PROPENSITY-SCORE MATCHING

Diqiong Xie, University of Iowa*
Michael P. Jones, University of Iowa

Propensity-score matching estimators for multiplicative treatment effects are widely used in observational studies despite the fact that their large sample properties have not been established. We derive the matching estimators and their large sample properties for the treatment effects that are measured as mean ratios and log mean ratios. Our theoretical results are developed in the framework of potential outcomes and conditional on matching with replacement. We propose and emphasize the use of matching calipers, the size of which varies with the number of matching candidates and the number of covariates involved in the propensity score model. We provide estimators for the large sample variances of matching estimators with a changing number of matches. A simulation study is conducted to support our theoretical findings. Various matching methods, i.e. 1:1 and M:1 matching with and without replacement and with and without calipers, are compared in the simulation study.

email: diqiong-xie@uiowa.edu

PRINCIPAL STRATIFICATION FOR ASSESSING MEDIATION WITH A CONTINUOUS MEDIATOR

Robert Gallop, West Chester University*

In assessing the mechanism of treatment efficacy in randomized clinical trials, investigators often perform mediation analyses by analyzing if the significant intent-to-treat treatment effect on outcome occurs through or around a third intermediate or mediating variable: indirect and direct effects, respectively. Standard mediation analyses assume sequential ignorability, i.e., conditional on covariates the intermediate or mediating factor is randomly assigned, as is the treatment in a randomized clinical trial. This research focuses on the application of the principal stratification approach for estimating the direct effect of a randomized treatment. Previous research ion this topic focused on a binary mediator, where the direct effect of treatment is estimated as a difference between expectations of potential outcomes within latent sub-groups, determined as a function of the two levels of the mediator, of participants for whom the intermediate variable behavior would be constant, regardless of the randomized treatment assignment. This current research extends this modeling structure to accommodate a continuous mediator. Using a Bayesian estimation procedure, we will estimate the direct effect of treatment and illustrate a direct effect curve to summarize the continuous mediator's mediation impact.

email: rgallop@wcupa.edu

TARGETED MINIMUM LOSS BASED ESTIMATION OF CAUSAL EFFECTS OF MULTIPLE TIME POINT INTERVENTIONS

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

Causal effect estimation in high-dimensional longitudinal data must appropriately account for time-dependent confounding, yet model mis-specification at multiple time steps can amplify bias. For this reason semi-parametric methods have an advantage over their parametric counterparts, and double robust estimators are

preferred. In this talk we show how the framework of targeted minimum loss based estimation can incorporate key ideas from the double robust estimating equation method proposed in Bang and Robins (2005) to produce a TMLE that 1) incorporates data adaptive estimation in place of parametric models, 2) can be applied to parameters for which there exists no mapping of the efficient influence curve into an estimating equation, thus also avoiding the potential problem of estimating equations having no or multiple solutions, and 3) has flexibility to incorporate robust choices of loss functions and hardest parametric sub-models so that the resulting TMLE is a robust substitution estimator. This new TMLE can be applied to survival analysis, estimation of time-dependent treatment effects, and generalizes to causal parameters defined by projections on working marginal structural models.

email: sgruber@hsph.harvard.edu

A DATA-ADAPTIVE APPROACH FOR MODELING PROPENSITY SCORES

Yeying Zhu, The Pennsylvania State University*
Debashis Ghosh, The Pennsylvania State University
Nandita Mitra, University of Pennsylvania
Bhramar Mukherjee, University of Michigan

In non-randomized observational studies, estimated differences between treatment groups may arise not only due to the treatment but also because of the masking effect of confounders. Therefore, causal inference regarding the treatment effect is not as straightforward as in a randomized trial. To adjust for confounding due to measured covariates, the average treatment effect is often estimated conditioning on propensity scores. Typically, propensity scores are estimated by logistic regression. Alternatively, one can employ nonparametric classification algorithms, such as tree-based approaches or support vector machines. In this talk, we explore the effect of classification algorithms used to model propensity scores using ideas of bias and variance. In addition, we explore ways to combine parametric models with nonparametric approaches to estimate propensity scores. Simulation studies are used to assess the performance of the newly proposed methods, and a data analysis example from the Surveillance, Epidemiology and End Results (SEER) database is presented.

email: yxz165@psu.edu

33. EPIDEMIOLOGIC METHODS

SEMI-PARAMETRIC METHODS FOR RELATIVE RISK CENTER EFFECT MEASURES

Kevin He, University of Michigan*
Douglas E. Schaubel, University of Michigan

The additive and multiplicative hazards models provide two frequently used frameworks for the analysis of survival data. We develop methods for evaluating center-specific survival using a center-stratified additive hazards model. The major difference between this model and the commonly used survival models is that it allows the regression effect to be additive, in the meanwhile it also allow the baseline hazards be center-specific. We then estimate the relative center effects by the ratio of survival functions. The proposed measure is a semiparametric generalization of the relative risk, which is often used in clinical studies. One advantage for our proposed method is that the ratio of survival function for a particular subject reduces to the ratio of baseline survival function, and such ratio of baseline survival functions is invariant to the choice of baseline covariate level. Therefore, the ratio of survival functions represents the contrast between subject i at center j versus subject i at the hypothetical center with baseline hazard function equal to the national average; where subject i can have any covariate value. We derive the asymptotic properties of the proposed estimators, and assess finite-sample characteristics through simulation. The proposed method is applied to national kidney transplant data.

email: kevinhe@umich.edu

A GENERAL BINOMIAL REGRESSION MODEL FOR ESTIMATING STANDARDIZED RISK DIFFERENCES FROM COHORT DATA

Stephanie A. Kovalchik, National Cancer Institute, National Institutes of Health*
Ravi Varadhan, Johns Hopkins University School of Medicine
Barbara Fetterman, Kaiser Permanente
Nancy E. Poitras, Kaiser Permanente
Sholom Wacholder, National Cancer Institute, National Institutes of Health
Hormuzd A. Katki, National Cancer Institute, National Institutes of Health

Absolute risk differences are of central importance to epidemiology and public health, yet reliable, easy-to-use methods to estimate standardized risk differences from binomial data are lacking. The linear-expit model (LEXPIT) is a flexible regression model for binary cohort data that combines linear and nonlinear effects, where the nonlinear term is the inverse-logit function and the linear coefficients are confounder-adjusted risk differences. For complex risk association studies, the LEXPIT model is advantageous because confounders are allowed to have linear or multiplicative

effects on disease risk. The LEXPIT parameters are estimated with a constrained maximum likelihood algorithm that ensures that estimated risks are probability measures. I will give an overview of the LEXPIT methodology and demonstrate its implementation using the R package blm in an application study to estimate absolute risk of cervical precancer or cancer for different Pap and human papillomavirus test results from 167,171 women at Kaiser Permanente Northern California.

email: kovalchiksa@mail.nih.gov

prLOGISTIC: AN R PACKAGE FOR ESTIMATION OF PREVALENCE RATIOS USING LOGISTIC MODELS

Leila D. Amorim, University of North Carolina at Chapel Hill*
Raydonal Ospina, Federal University of Pernambuco, Brazil

The interpretation of odds ratios (OR) as prevalence ratios (PR) in cross-sectional studies has been criticized since this equivalence is not true unless under specific circumstances. The logistic model is a very well known statistical tool for analysis of binary outcomes and frequently used to obtain adjusted OR. Several statistical models are discussed in the literature to provide adjusted estimates of PR, including logistic model, Poisson regression and log-binomial regression. Since logistic regression is the most popular model for analysis of binary outcomes, its use is appealing for estimation of PR. Another issue that has been discussed in the literature is the estimation of adjusted PR for correlated data (Santos et al, 2008). We describe the R package prLogistic for estimation of PR using logistic models for analysis of independent and correlated binary data. We show how to use the package, considering two standardization procedures. Delta method and bootstrap are used for obtaining confidence intervals for PR. Our package includes several datasets used to illustrate its application for analysis of independent observations and clustered studies.

email: lamorim@email.unc.edu

EXTENDING MATRIX AND INVERSE MATRIX METHODS: ANOTHER LOOK AT BARRON'S APPROACH

Li Tang, Emory University*
Robert H. Lyles, Emory University
David D. Celantano, Johns Hopkins Bloomberg School of Public Health
Yungtai Lo, Montefiore Medical Center and Albert Einstein College of Medicine

The problems of misclassification are common in epidemiological and clinical research. Sometimes misclassification may exist in both exposure and outcome variables. It is well known that validity of analytic results (e.g., estimates of odds ratios of interest) might be questionable when no correction effort is made. Therefore, valid and accessible methods with which to deal with these issues are still in high demand. Here we elucidate extensions of well-studied methods in order to facilitate misclassification adjustment when a binary outcome and binary exposure variable are both subject to misclassification. By formulating generalizations of assumptions underlying Barron's original matrix method and the original inverse matrix method into the framework of maximum likelihood, our approach allows the incorporation of covariates both in the main health effects model of interest and in misclassification models for the binary outcome and exposure variable. We illustrate how the approach can adjust for differential misclassification in both variables when adequate internal validation data are available. The value of our extensions is demonstrated by means of simulations, and (as time permits) by means of a motivating example.

email: ltang3@emory.edu

PATTERN-MIXTURE MODELS FOR ADDRESSING OUTCOME MISCLASSIFICATION FROM PROXIES RESPONDING ON BEHALF OF PARTICIPANTS WITH INFORMATIVELY MISSING SELF-REPORTS

Michelle Shardell, University of Maryland School of Medicine*

Proxy respondents, such as relatives or caregivers, are often recruited in epidemiological studies of older adults when study participants are unable to provide self-reports (e.g., due to cognitive impairment). For each participant, either a proxy or participant response, but not both, is available for analysis. Substituting proxy responses for missing participant data introduces misclassification error and leads to biased parameter estimates. However, the mechanism for missing participant data is unknown and cannot be identified by the data. We propose a pattern-mixture model to avoid bias from missing participant data while leveraging the proxy data. For observations with missing participant responses, we use exponential tilt models to relate non-identifiable covariate-specific sensitivity and specificity parameters to estimable proxy outcome proportions while satisfying restrictions on sensitivity and specificity. We accommodate high-dimensional covariates and circumvent model incompatibility via propensity score stratification. Simulation studies show that the proposed method performs well and has low bias. The method is applied a cohort of elderly hip fracture patients.

email: mshardel@epi.umaryland.edu

34. RECENT ADVANCES ON HIGH-DIMENSIONAL MEDICAL DATA ANALYSIS

FEATURE SCREENING VIA DISTANCE CORRELATION LEARNING

Runze Li, Penn State University*
Wei Zhong, Penn State University
Liping Zhu, Shanghai University of Finance and Economics

This paper is concerned with screening features in ultrahigh dimensional data analysis, which has become increasingly important in diverse scientific fields. We develop a sure independence screening procedure based on the distance correlation (DC-SIS, for short). The DC-SIS can be implemented as easily as the sure independence screening procedure based on the Pearson correlation (SIS, for short) proposed by Fan and Lv (2008). However, the DC-SIS can significantly improve the SIS. Fan and Lv (2008) established the sure screening property for the SIS based on linear models, but the sure screening property is valid for the DC-SIS under more general settings including linear models. Furthermore, the implementation of the DC-SIS does not require model specification (e.g., linear model or generalized linear model) for responses or predictors. This is a very appealing property in ultrahigh dimensional data analysis. Moreover, the DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables. We study the theoretic properties of the DC-SIS, establish its ranking consistency and sure screening properties, and conduct simulations to examine its finite sample performance. Numerical comparison indicates that the DC-SIS performs much better than the SIS in various models. We illustrate the DC-SIS through two real data examples.

email: rli@stat.psu.edu

TIME-VARYING SIGNAL DETECTION FOR CORRELATED DATA

Annie Qu, University of Illinois at Urbana-Champaign*
Lan Xue, Oregon State University
Colin Wu, National Heart, Lung and Blood Institute, National Institutes of Health

This talk is motivated by the National Heart, Lung and Blood Institute Growth and Health Study (NGHS) which evaluates the longitudinal effects of race, height and body-mass index on the levels of systolic blood pressure and diastolic blood pressure. Here the signals associated with relevant predictors could be time-dependent. For NGHS data, it has been observed that children's and adolescents' heights seem to be associated with high blood-pressure during certain age periods. Therefore it is also scientifically important to detect relevant covariate effects which could be time-dependent. We propose a varying-coefficient model selection and estimation which can capture relevant time-dependent covariates. Because of nonparametric components are involved for the varying-coefficient model, we are dealing with a high-dimensional parameters problem. The proposed method will be illustrated using the NGHS data.

email: anniequ@illinois.edu

SOFARE: SELECTION OF FIXED AND RANDOM EFFECTS IN HIGH-DIMENSIONAL LONGITUDINAL DATA ANALYSIS

Yun Li, University of Michigan
Sijian Wang, University of Wisconsin-Madison
Peter X.K. Song, University of Michigan*
Naisyin Wang, University of Michigan
Ji Zhu, University of Michigan

SOFARE is a new algorithm of variable selection in regularized mixed-effects regression models for high-dimensional longitudinal data. The proposed regularization takes place simultaneously at both fixed effects and random effects, in which estimation and selection in the mean and covariance structures are carried out via penalized likelihood and penalized REML, respectively, under scenarios of P>>n and P <n. An application of SOFARE is to detect any predictors that are nonlinearly associated with outcomes through semiparametric additive mixed-effects models. SOFARE enables us to automatically determine which predictors are unassociated, linearly associated or nonlinearly associated with outcomes. We demonstrate SOFARE using both simulation studies and a real-world data example of longitudinal cohort study on diabetic nephropathy.

email: pxsong@umich.edu

VARIABLE SELECTION FOR OPTIMAL TREATMENT DECISION

Hao Helen Zhang, University of Arizona*
Wenbin Lu, North Carolina State University
Donglin Zeng, University of North Carolina at Chapel Hill

In decision-making on optimal treatment strategies, it is of great importance to identify variables that are involved in the decision rule, i.e. those interacting with the treatment. Effective variable selection helps to improve the prediction accuracy and enhance the interpretability of the decision rule, especially for high dimensional data. We propose a new regression framework to simultaneously estimate the optimal treatment strategy and identify important variables. The new approach does not require estimation of the baseline mean function of the response and hence greatly improves the robustness of the estimator. Its convenient loss function makes it easy to adopt modern shrinkage methods for variable selection, which facilitates implementation and statistical inferences for the estimator. The new procedure can be easily implemented by existing software packages. Theoretical properties are studied. Its empirical performance is evaluated using simulation studies and illustrated with an application to an AIDS clinical trial.

email: hzhang@math.arizona.edu

35. BAYESIAN APPROACHES WITH APPLICATIONS TO GENOMICS

BAYESIAN INFERENCE OF CHROMOSOME LOCAL 3D STRUCTURES FROM HI-C DATA

Ming Hu, Harvard University*
Ke Deng, Harvard University
Zhaohui S. Qin, Emory University
Jun S. Liu, Harvard University

How chromatin fits into a nucleus remains largely unresolved. Understanding how chromosomes fold provides insights into transcription regulation hence functional state of the cell. Recently, Hi-C technology has been developed to provide an unbiased view of chromatin organization in the nucleus. In the Hi-C experiment, the spatial proximity of any two genomic loci is represented by a count matrix. The goal of our study is to translate the count matrix into three-dimensional (3D) structure of local genomic domains. To achieve this, we devise a novel Bayesian statistical model linking the observed read counts spanning a pair of loci to the spatial distance between them. Using advanced Monte Carlo computational techniques such as sequential Monte Carlo and hybrid Monte Carlo, we are able to reconstruct the spatial arrangement of local genomic domains in 3D space. When applying our method to a real Hi-C dataset, we can visualize the 3D shape of the local genomic domains, and the predicted spatial distances are consistent with the gold standard florescent in situ hybridization (FISH) data. With the explosive accumulation of high throughput genome-wide chromatin interaction data, the proposed method will have immediate and far-reaching impact in the broader area of biomedical research.

email: minghu@fas.harvard.edu

INFERRING SOCIAL NETWORKS FROM MOLECULAR AND LINGUISTIC DATA

Marc A. Suchard, University of California at Los Angeles*

Phylogeography attempts to connect the evolutionary and spatial histories of biologically-related organisms. Previous approaches have assumed that the spatial contact network is known. In this talk, I describe recent Bayesian extensions to simultaneously infer the unknown relatedness between organisms and determinants of the unknown spatial structure. Simultaneous inference accounts for uncertainty in both and allows for the injection of prior information. Bayesian non-parametric models and high-performance computing exploiting massive parallelization make this possible. To demonstrate these novel methods, I explore the phylogeographic histories of influenza A viruses, determine the most probable origins of Avian and Swine flu and examine the unknown contact network in seasonal epidemics. Air-flight patterns fall out as the most significant predictor of this network. I finish with a novel application of Bayesian phylogeography to human pre-history. I examine the origins of Indo-European speaking people, doubted by Jared Diamond "most recalcitrant problem in historical linguistics," and find strong statistical support for an Anatolian, as opposed to a Pontic steppes, homeland.

email: msuchard@ucla.edu

BAYESIAN APPROACHES FOR THE INTEGRATION OF LARGE-SCALE DATA

Marina Vannucci, Rice University*

Novel methodological questions are now being generated in Bioinformatics and require the integration of different concepts, methods, tools and data types. In this talk I will explore Bayesian graphical modeling approaches that infer biological regulatory networks by integrating expression levels of different types. The proposed modeling strategy is general and can be easily applied to different types of network inference. In one example I will consider models that relate miRNA expression data to mRNAs. I will also look at models that relate genotype data to mRNAs for the selection of the markers that affect the gene expression. Specific sequence/structure information will be incorporated into the prior probability models.

email: marina@rice.edu

BAYESIAN HIERARCHICAL GRAPH-STRUCTURED MODEL WITH APPLICATION TO PATHWAY ANALYSIS USING GENE EXPRESSION DATA

Hui Zhou, Columbia University
Tian Zheng, Columbia University*

Graph constrained (or structure) inference takes advantage of a known relational structure among variables to introduce smoothness and reduce complexity in modeling, especially for high-dimensional data such as that in genomics. There has been a lot of interest in its application in model regularization and selection. However, prior knowledge on the graphical structure among the variables can be limited and partial. Empirical data may suggest modifications to such a graph, which could lead to new and interesting biological findings. In this paper, we propose a Bayesian random graph structured model, rGrace, an extension from the Grace model by Li and Li (2010), to combine a priori network information and empirical evidence, for applications such as pathway analysis. Using both simulations and real data examples, we show that the new method can identify discrepancy between data and a prior known graph structure and suggest modifications and updates.

email: tz33@columbia.edu

36. NEW TRENDS IN STATISTICAL ANALYSIS OF BIOLOGICAL NETWORKS

NEW TOOLS FOR SYSTEMS-LEVEL ANALYSIS OF REGULATION AND SIGNALING DYNAMICS

Alexander Franks, Harvard University
Edoardo M. Airoldi, Harvard University*

Mapping the functional landscape driving complex cellular phenotypes is a central goal of modern genome- and proteome-scale studies. In this talk, I will present new models that support such analysis. First, we will consider perturbation experimental designs where a quantitative trait of interest Y, such as gene expression is measured multiple times, for different values of a covariate X, such as time or growth rate, and for different factors F, such as knock-out strains or growth media. In this context, I will introduce a linear model framework where the linear response of individual genes depends explicitly on the linear response of the functions these genes are annotated to, according to the gene ontology. We will use this model to study the functional basis of the cellular response to a series of nutrient perturbations, in yeast. Second, we will consider coordinated experimental designs where we observe multiple high-dimensional phenotypical responses Y1...Yk over time. In this context, we want to quantify the extent to which the newly generated data supports current hypotheses about regulation and signaling dynamics, and we want to generate novel hypotheses that can be tested at the bench. We will use this model to explore the pheromone response pathway, in yeast.

email: airoldi@fas.harvard.edu

DYNAMIC MODELS FOR BABOON GROOMING NETWORKS

David L. Banks, Duke University*
Yingbo Li, Duke University

Baboon social networks are largely expressed through grooming relations. This work analyzes three years of field data from the Amboseli National Reserve in Kenya, in order to model the fission of a baboon trip into two smaller groups. The analysis uses covariates on kinship, genetics, social structure, and rainfall, together with a dynamic model that incorporates reciprocity and triad completion, to describe the changing structure of grooming relations in the community.

email: banks@stat.duke.edu

BIOLOGICALLY-STRUCTURED LATENT FACTOR MODELS FOR IDENTIFICATION OF CELLULAR MECHANISM OF ACTION

Lisa Pham, Boston University
Eric D. Kolaczyk, Boston University*
Luis E. Carvalho, Boston University
Stephane Robin, ParisAgroTech
Scott E. Schaus, Boston University

Identifying biological mechanisms of action (e.g. biological pathways) that control disease states, drug response, and altered cellular function is a multifaceted problem involving a dynamic system of biological variables that culminate in an altered cellular state. The challenge is in deciphering the factors that play key roles in determining the cells fate. We describe a modeling framework that addresses this problem, using gene expression data in conjunction with information from biological databases. More specifically, our framework models gene expression as a function of a perturbed latent biologically-informed, inter-connected network of pathways. The underlying goal in this setting is to identify the primary perturbed biological pathways of a given experiment, whose effects propagate through the rest of the pathway network, ultimately affecting the observed gene expression. We conduct appropriate inference using an MCMC algorithm. Although the algorithm admits some parallelization, it is still computationally demanding in practice. We therefore explore a variational Bayes approach as well. Simulation results are presented comparing these two approaches. We illustrate using gene transcription cancer profiles from The Cancer Genome Atlas Database.

email: kolaczyk@bu.edu

INFERRING GENE REGULATORY NETWORKS BY INTEGRATING PERTURBATION SCREENS AND STEADY-STATE EXPRESSION PROFILES

Ali Shojaie, University of Washington*
Alexandra Jauhiainen, University of Michigan
Michael Kallitsis, University of Michigan
George Michailidis, University of Michigan

Reconstructing transcriptional regulatory networks is an important task in systems biology. Data obtained from experiments that perturb genes by knock-outs or RNA interference contain useful information for addressing the reconstruction problem. However, such data can be limited in size and/or expensive to acquire. On the other hand, observational data of the organism in steady state are more readily available, but their informational content inadequate for the task at hand. We develop a computational approach to appropriately utilize both data sources for estimating a regulatory network, using a three-step algorithm that uses as input both perturbation screens and steady state gene expression data. In the first step, the algorithm determines causal orderings of the genes that are consistent with the perturbation data. In the second step, for each ordering, a regulatory network is estimated using a penalized likelihood based method, while in the third step

a consensus network is constructed from the highest scored ones. Further, it is established that the algorithm produces a consistent estimate of the regulatory network, and allows for existence of cycles in the network. Numerical results show that the algorithm performs well in uncovering the underlying network and clearly outperforms competing approaches that rely only on a single data source.

email: ashojaie@uw.edu

37. MATHEMATICAL MODELING OF DISEASE

DYNAMICS OF TREATMENT RESPONSES IN CHRONIC MYELOID LEUKAEMIA

Min Tang, Dana-Farber Cancer Institute and Harvard School of Public Health*
Franziska Michor, Dana-Farber Cancer Institute and Harvard School of Public Health
Mithat Gonen, Memorial Sloan-Kettering Cancer Center
Alfonso Quintas-Cardama, University of Texas MD Anderson Cancer Center
Jorge Cortes, University of Texas MD Anderson Cancer Center
Hagop Kantarjian, University of Texas MD Anderson Cancer Center
Chani Field, University of Adelaide, Adelaide, Australia
Timothy P. Hughes, University of Adelaide, Adelaide, Australia
Susan Branford, University of Adelaide, Adelaide, Australia

Chronic Myeloid Leukaemia (CML) represents the first human cancer in which molecularly targeted therapy leads to a dramatic clinical response. Imatinib mesylate (Gleevec) and nilotinib are potent inhibitors of the BCR-ABL fusion oncogene that drives the leukemia. Although CML represents one of the most well-studied cancers, several critical questions remain such as the treatment response of CML stem cells. In order to investigate the behavior of leukemic stem cells during treatment, we analyzed the long-term IRIS trial data, where 29 newly diagnosed CML patients were treated with first-line imatinib for up to 10 years. We utilized a statistical modeling approach to identify the shape of the treatment response curves in this cohort as well as in a short-term IRIS trial cohort, which has better data resolution to refer the treatment kinetics of initial decline. We found that the imatinib treatment response of BCR-ABL1 transcripts in the peripheral blood of most patients displays three phases. Together with a four-compartment mathematical framework, which can explain the kinetics of the molecular response to treatment therapy, we concluded that targeted therapy is capable of depleting an immature leukemic cell population, possibly leukemic stem cells, at a very slow rate in a subset of patients.

email: min@jimmy.harvard.edu

MATHEMATICAL MODELING OF PANCREATIC CANCER PROGRESSION REVEALS DYNAMICS OF GROWTH AND DISSEMINATION AND SUGGESTS OPTIMUM TREATMENT STRATEGIES

Hiroshi Haeno, Dana-Farber Cancer Institute
Mithat Gonen, Memorial Sloan-Kettering Cancer Center
Meghan Davis, Johns Hopkins University
Joseph Herman, Johns Hopkins University
Christine Iacobuzio-Donahue, Johns Hopkins University
Franziska Michor, Dana-Farber Cancer Institute*

Pancreatic cancer is a leading cause of cancer-related death, largely due to metastatic dissemination. We investigated pancreatic cancer progression by utilizing a mathematical framework of metastasis formation together with comprehensive data of 228 patients, 101 of whom had autopsies. We found that pancreatic cancer growth is initially exponential; however, primary and metastatic sites exhibit independent growth kinetics. After estimating the rates of pancreatic cancer progression and dissemination, we determined that patients likely harbor metastases at diagnosis and predicted the number and size distribution of metastases as well as patient survival. These findings were validated in an independent database. Finally, we analyzed the effects of different treatment modalities, finding that therapies which efficiently reduce the growth rate of cells earlier in the course of treatment appear to be superior to upfront tumor resection. This interdisciplinary approach provides insights into the dynamics of pancreatic cancer metastasis and identifies optimum therapeutic interventions.

email: michor@jimmy.harvard.edu

PATIENT-SPECIFIC MATHEMATICAL MODELING OF GLIOMA PROLIFERATION AND INVASION: INFORMING TREATMENT DESIGN AND PATIENT STRATIFICATION

Kristin Swanson, University of Washington*
Russ Rockne, University of Washington
Dave M. Corwin, University of Washington
Robert Stewart, University of Washington
Mark Philips, University of Washington
Clay Holdsworth, University of Washington
Andrew Trister, University of Washington
Jason Rockhill, University of Washington
Maciej Mrugala, University of Washington

Glioblastomas are primary brain tumors known for their diffuse invasion beyond imaging margins and their heterogeneity within and across patients. This challenges the optimal design of therapies (surgical, radio- and chemo- therapy) in individual patients. To this end, using an iterative dialog between our patient-specific mathematical model for brain tumor growth which quantifies net rates of proliferation and invasion and radiation sensitivity [Rockne 2010], and routine clinical imaging followup, we have developed a means of quantifying and predicting tumor growth and progression in individual patients. This provides a novel opportunity for patient-individualized treatment design and patient stratification in clinical studies to determine the optimal

therapies for each patient. The ultimate clinical relevance of this work lies in its potential to shape the future of personalized medicine through optimized treatment plans targeted at the individual patient's disease kinetics.

email: krae@uw.edu

38. HIGH DIMENSIONAL MULTI-DRUG COMBINATIONS: FROM PRECLINICAL MODELS TO CLINICAL TRIALS

STATISTICAL METHODS FOR PRECLINICAL MULTI-DRUG COMBINATION

Ming T. Tan, University of Maryland School of Medicine*

Drug combinations are the hallmark of cancer therapy and are used widely in other complex diseases such as hypertension and infectious diseases as well. Preclinical experiments on multi-drug combinations are important steps to bring the therapy to clinic. A statistical approach for evaluating the joint effect of the combination is necessary because even in vitro experiments often demonstrate significant variation in dose-effect. Such variation needs to be accounted for in the experimental design and analysis. We present a novel methodology for experimental design and analysis of preclinical drug combination studies. We then focus on the design and analysis of two new anti-cancer drug development experiments. We will highlight the contributions of the novel statistical methodology brought to the therapeutic development of the two drug combination therapies. We demonstrate that these statistical methods and software have resulted in the identification of highly synergistic dose combinations that could have been missed with classic methods. This work is supported in part by NCI and in collaboration with Hongbin Fang, Doug Ross and Martin Edelman.

email: mttan@som.umaryland.edu

DOSE-FINDING METHODS FOR COMBINATIONS OF AGENTS

Mark R. Conaway, University of Virginia*

Dose-finding or phase I trials of treatments that are combinations of agents are becomingly increasingly common in oncology research. These studies attempt to identify a treatment that can be administered with an acceptable level of toxicity. In many cases, standard designs for phase I trials are not appropriate because these designs are based on the assumption that the treatments can be ordered, before the study begins, with respect to the probability of toxicity. This talk will present a number of methods, including those of Wages, Conaway and O'Quigley (2011) for the design of dose-finding trials of combinations of agents.

email: mconaway@virginia.edu

A BAYESIAN DOSE-FINDING DESIGN FOR DRUG COMBINATION TRIALS WITH DELAYED TOXICITIES

Suyu Liu, University of Texas MD Anderson Cancer Center
Ying Yuan, University of Texas MD Anderson Cancer Center*

We propose a Bayesian adaptive dose-finding design for combination trials with delayed outcomes. We model the dose-toxicity relationship using the Finney model, a drug-drug interaction model that has been validated by many empirical studies and extensively investigated in the drug-drug interaction literature. The parameters in the Finney model has intuitive interpretations, which greatly facilitates incorporating the available prior dose-toxicity information from single-agent trials through prior elicitation. To accommodate delayed outcomes, we treat unobserved delayed toxicity outcomes as missing data and use Bayesian data augment to handle the resulting missing data. We conduct extensive simulation studies to examine the operating characteristics of the proposed method under various practical scenarios. Results show that the proposed design is safe and able to select target dose combinations with high probabilities. The proposed design also satisfactorily addresses the delayed outcomes and allows a fast continuous accrual.

email: yyuan@mdanderson.org

39. GROUP TESTING METHODOLOGY: RECENT DEVELOPMENTS AND APPLICATIONS TO INFECTIOUS DISEASE

MARGINAL REGRESSION MODELS FOR MULTIPLE-DISEASE GROUP TESTING DATA

Christopher R. Bilder, University of Nebraska-Lincoln*
Boan Zhang, University of Nebraska-Lincoln
Joshua M. Tebbs, University of South Carolina

Group testing, where groups of individual specimens are composited to test for the presence of a disease (or other binary trait), is a procedure commonly used to reduce the costs of screening a large number of individuals. Group testing data are unique in that only group responses may be observed, but inferences are needed at the individual level. A further methodological challenge arises when individuals are tested in groups for multiple diseases simultaneously, because the unobserved individual disease statuses are likely to be correlated. In this paper, we propose the first regression techniques for multiple-disease group testing data. We develop an expectation-solution based algorithm that provides consistent parameter estimates and natural large-sample inference procedures. Our proposed methodology is applied to chlamydia and gonorrhea screening data collected in Nebraska as part of the Infertility Prevention Project.

email: cbilder3@unl.edu

SYSTEM OF EQUATIONS APPROACH TO POOLED NUCLEIC ACID TESTING FOR FAILING ANTIRETROVIRAL THERAPY

Tanya S. Granston, University of Washington*
Susanne May, University of Washington
Davey M. Smith, University of California at San Diego

Periodic individual monitoring of HIV viral loads for identifying failed anti-retroviral therapy (ART) is financially prohibitive in resource-limited settings. Pooling strategies like the matrix approach have been shown to reduce the number of tests necessary to distinguish the failures when the prevalence of ART failure is between 1% and 25%, and with relatively high negative predictive values. However, maximum relative efficiencies are only a little more than 30% (50%) relative to individual testing when ART is failing in 20% (10%) of the cohort. We expand on the matrix approach and view the unknown matrix and the pooling information as a system of equations. There is no unique solution to this system of equations, however, under certain distributional assumptions, the optimal solution that resolves the matrix can be selected from a set of feasible permutations of solutions. In some scenarios, this approach can be more efficient in terms of cost and turn-around time and may be considered superior to the basic matrix approach to pooled testing.

email: granston@uw.edu

TWO-DIMENSIONAL INFORMATIVE ARRAY TESTING

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

Array-based group testing algorithms for case identification have been widely considered for use in infectious disease testing, drug discovery, and genetics. In this paper, we generalize previous statistical work in array testing to account for heterogeneity among individuals being tested. We first derive closed-form expressions for the expected number of tests (efficiency) and misclassification probabilities (sensitivity, specificity, predictive values) for two-dimensional array testing in a heterogeneous population. We then propose two 'informative' array construction techniques which exploit population heterogeneity in ways that can substantially improve testing efficiency when compared to classical approaches which regard the population as homogeneous. Furthermore, a useful byproduct of our methodology is that misclassification probabilities can be estimated on a per-individual basis. We illustrate our new procedures using chlamydia and gonorrhea testing data collected in Nebraska as part of the Infertility Prevention Project.

email: tebbs@stat.sc.edu

40. NOVEL DEVELOPMENTS IN STATISTICAL BLIND SOURCE SEPARATION AND INDEPENDENT COMPONENTS ANALYSIS

A NEW PROBABILISTIC GROUP ICA METHOD FOR MODELING BETWEEN-SUBJECT VARIABILITY IN BRAIN FUNCTIONAL NETWORKS

Ying Guo, Emory University*
Li Tang, Emory University

Independent component analysis (ICA) has become an important tool for identifying and characterizing spatial distributed patterns of brain functional networks. The original ICA algorithm was developed for single-subject analysis. Several group ICA methods have been proposed for multi-subject data. A major limitation of these methods is that they typically assume common spatially distributed patterns across subjects in ICA decomposition. Between-subject variability is currently addressed in post-ICA analysis through heuristically reconstructed subject-specific spatial maps but is not directly accounted for or characterized in group ICA. I propose a new probabilistic group ICA model that directly models subject-specific effects in ICA decomposition, therefore providing more accurate statistical inference for brain functional networks on both the population and subject level.

email: yguo2@emory.edu

NONPARAMETRIC INDEPENDENT COMPONENT ANALYSIS WITH APPLICATION TO EEG DATA

Seonjoo Lee, Henry Jackson Foundation*
Haipeng Shen, University of North Carolina at Chapel Hill
Young Truong, University of North Carolina at Chapel Hill

Independent component analysis (ICA) is an effective data-driven method for blind source separation. It has been successfully applied to separate source signals of interest from their mixtures. Most existing ICA procedures are for instantaneous mixtures and carried out by relying solely on the estimation of the marginal density functions. In many practical applications, the sources have temporal autocorrelations, or are mixed with time delays. For the convolutive mixtures, convolutive ICA based on ARMA models have been proposed. In this talk, we consider the case of sources with possibly mixed spectra, where ARMA estimates are often unstable. Specifically, we propose to estimate the spectral density functions and line spectra of the source signals using cubic splines and indicator functions, respectively. The mixed spectra and the mixing matrix are estimated via maximizing the Whittle likelihood function. We illustrate the performance of the proposed method through extensive simulation studies and a resting stage EEG data application. The numerical results indicate that our approach outperforms existing ICA methods including the most widely used Infomax algorithm.

email: seonjool@gmail.com

INDEPENDENT COMPONENT ANALYSIS FOR FUNCTIONAL IMAGING DATA

Ani Eloyan, Johns Hopkins University*
Brian Caffo, Johns Hopkins University
Ciprian Crainiceanu, Johns Hopkins University

With the development of functional imaging technology large amounts of imaging data are available for constructing automated methods for disease diagnostics, biomarker identification, etc. Independent component analysis (ICA) is a statistical method widely used in neuroscience for finding meaningful patterns in the data such as identifying functional networks in the brain using functional magnetic resonance imaging (fMRI) data. Several methods have been developed for ICA analysis of group imaging data. However, most of these methods are infeasible when the number of subjects in the study is large. We propose a likelihood based, iterative ICA method for group imaging data analysis which does not require loading the data for all subjects simultaneously, hence making the method applicable for many subjects. The algorithm also provides the densities of the underlying independent components as a byproduct. The method is applied to simulated data to show the empirical behavior. Further, a real functional data analysis example is presented.

email: aeloyan@jhsph.edu

INDEPENDENT COMPONENT ANALYSIS VIA DISTANCE COVARIANCE

David S. Matteson*, Cornell University
Ruey S. Tsay, University of Chicago

We introduce a novel statistical framework for independent component analysis (ICA) of multivariate data. We propose procedures for estimating and testing the existence of independent components for a given dataset. Independent components are estimated by combining a nonparametric probability integral transformation with a generalized nonparametric whitening method that simultaneously minimizes all forms of dependence among the components. The distance covariance statistics are combined in succession to construct a test for the existence of independent components. When independent components exist, one can apply univariate analysis to study and build a model for each component. Those univariate models are then combined to obtain a multivariate model for the original observations. We prove the consistency of our estimator under minimal regularity conditions. We demonstrate the improvements of the proposed method over competing methods in simulation studies. We apply the proposed ICA to two real examples and contrast it with principal component analysis.

email: dm484@cornell.edu

A BAYESIAN RANDOM SHAPE MODEL FOR fMRI AND MRI DATA

Lijun Zhang*, Emory University
Jian Kang, Emory University
F. DuBois Bowman, Emory University

Functional magnetic resonance imaging (fMRI) and MRI data provide us with noninvasive tools to better understand cognitive processes associated with specific brain regions and brain anatomy. They also reveal important functional and structural properties that distinguish subgroups of subjects, e.g. a patient population from healthy control subjects. We propose a Bayesian random shape model to partition the brain voxels into a set of disjoint shapes that are best representative of the activation regions (or alternatively structural features). Our method can simultaneously segment the voxels into different activation shapes and avoids the introduction of user defined seed regions. It achieves more accurate mapping of the stimulus-induced activated brain regions and can simplify group comparisons or classification by using the activation shapes. We apply our methods to an fMRI study to compare the results generated by a general linear model (GLM) and also to evaluate the differences in functional shape features between cocaine addicted men and healthy control subjects.

email: l.zhang@emory.edu

41. CAUSAL INFERENCE AND SURVIVAL ANALYSIS

ESTIMATING THE AVERAGE TREATMENT EFFECT ON MEAN SURVIVAL TIME WHEN TREATMENT IS TIME-DEPENDENT AND CENSORING IS DEPENDENT

Douglas E. Schaubel, University of Michigan*
Qi Gong, University of Michigan

We propose methods for estimating the average difference in restricted mean survival time attributable to a time-dependent treatment. In the data structure of interest, the time until treatment is received and the pre-treatment death hazard are both heavily influenced by a longitudinal process. In addition, subjects may experience periods of treatment ineligibility. The pre-treatment death hazard is modeled using inverse weighted partly conditional methods, while the post-treatment hazard is handled through Cox regression. Subject-specific differences in pre- versus post-treatment survival are estimated, then averaged in order to estimate the average treatment effect among the treated. Asymtptotic properties of the proposed estimators are derived and evaluated in finite samples through simulation. The proposed methods are applied to liver failure data obtained from a national organ transplant registry.

email: deschau@umich.edu

MATCHING METHODS FOR OBTAINING SURVIVAL FUNCTIONS TO ESTIMATE THE EFFECT OF A TIME-DEPENDENT TREATMENT

Yun Li, University of Michigan*
Douglas E. Schaubel, University of Michigan

In observational studies of survival time featuring a binary time-dependent treatment, the hazard ratio (an instantaneous measure) is often used to represent the treatment effect. However, investigators are often more interested in the difference in survival functions, which is not straightforward to obtain. We propose semi-parametric methods to estimate the causal effect on the post-treatment survival among the treated. The objective is to compare post-treatment survival among patients who receive treatment, with the survival function that would have been observed in the absence of treatment. We use various matching techniques to select comparable patients for each treated patient. After the adjustment covariate distributions are balanced through matching, non-parametric methods can be employed to estimate the survival functions. We conduct matching by comparing a treated patient with his/her potential matches by the closeness of their risk scores that either base on the death hazard or time-dependent treatment propensity or both. The censoring distribution is accounted for by inverse probability of censoring weighting. The proposed methods are applied to examine the effect of transplantation on the post-transplant survival curves among end-stage renal disease patients.

email: yunlisph@umich.edu

OPTIMIZATION OF DYNAMIC TREATMENT REGIMES FOR RECURRENT DISEASES

Xuelin Huang, University of Texas MD Anderson Cancer Center*
Jing Ning, University of Texas MD Anderson Cancer Center

Multi-stage treatments for recurrent diseases, such as many types of cancer, are inevitably dynamic. That is, the choices of the next treatment depend on the patient's responses to previous therapies. Dynamic treatment regimes (DTRs) are routinely used in clinics, but rarely optimized. In this talk, we distinguish two types of outcomes: cumulative and non-cumulative. Current methods in the literature focus on the optimization of non-cumulative outcomes. We point out that, for cumulative outcomes, such as the survival time of patients with a recurrent disease, a readily manageable statistical method is available for the optimization of their DTRs. Sequential accelerated failure time (AFT) models on counter-factual survival times are developed for this purpose. Comparing with current methods, the proposed method does not need the specification of treatment selection models. This avoids the bias and convergence problems due to the mis-specification of treatment selection models. Under some Markov conditions, the proposed method does not suffer from the curse of dimensionality as the number of treatment stages increases, which is a serious problem for the traditional methods under the same conditions. Simulation and real studies show that the proposed method performs well and is useful in practical situations.

email: xlhuang@mdanderson.org

PREDICTION OF SURVIVAL AND VARIABLE IMPORTANCE IN MEDICAL INFORMATICS: TARGETED MAXIMUM LIKELIHOOD ESTIMATION (T-MLE) AND SUPERLEARNING APPLIED TO HIGH DIMENSIONAL LONGITUDINAL DATA TO PREDICT SURVIVAL TIMES AMONG SEVERE TRAUMA PATIENTS

Alan Hubbard, University of California, Berkeley*
Mitch Cohen, University of California, San Francisco
Anna Decker, University of California, Berkeley
Ivan Diaz, University of California, Berkeley
Matthew Kutcher, University of California, San Francisco

As complex high dimensional, longitudinal data becomes more routinely gathered on patients in hospital or clinical settings, so have statistical techniques with the emphasis to use such data to predict outcomes, and find explanatory patterns among such data. Often the underlying implied data-generated models from such procedures put arbitrary constraints on the underlying data-generating distribution, about which typically little is known (true model semiparametric). When machine learning algorithms have been used, researchers may interpret the implied importance of variables from a single fit via a data-adaptive procedure, and these importances have unpredictable sampling distributional properties. We discuss estimation in a semiparametric model of asymptotically linear parameters, that have desirable sampling-distributional properties. We define our parameters in the context of graphs and thus the variable importance as theoretical interventions on a set of implied nonparametric structural equation models. The approach used is targeted maximum likelihood estimation and superLearning and these are applied to develop time-dependent prognostic indicators of survival (and other time-to- event outcomes) among a group of severe trauma patients, based on high dimensional, time-dependent data.

email: hubbard@berkeley.edu

A SEMIPARAMETRIC RECURRENT EVENTS MODEL WITH TIME-VARYING COEFFICIENTS

Zhangsheng Yu, Indiana University School of Medicine*
Lei Liu, University of Virginia

We consider a recurrent events model with time-varying coefficients motivated by two clinical applications. A random effects (Gaussian frailty) model is used to describe the intensity of recurrent events. The model can accommodate both time-varying and constant coefficients. The penalized spline method is used to estimate the time-varying coefficient. Laplace approximation is used to evaluate the penalized likelihood without a closed form. The smoothing parameters are estimated in a similar way to variance components. We conduct simulations to evaluate the performance of the estimate for both time-varying and time-independent coefficients. We apply this method to analyze two data sets: a stroke study and a child wheeze study.

email: yuz@iupui.edu

42. CLINICAL TRIALS

ESTIMATING COVARIATE-ADJUSTED LOG HAZARD RATIOS IN RANDOMIZED CLINICAL TRIALS USING COX PROPORTIONAL HAZARDS MODELS AND NONPARAMETRIC RANDOMIZATION BASED ANALYSIS OF COVARIANCE

Benjamin R. Saville, Vanderbilt University*
Gary G. Koch, University of North Carolina at Chapel Hill

In the context of randomized clinical trials with time-to-event outcomes, estimates of covariate-adjusted log hazard ratios for comparing two treatments are obtained via nonparametric analysis of covariance by forcing the difference in means for covariables to zero. The method avoids the assumption of proportional hazards for each of the covariates, and it provides an adjusted analysis for the same population average treatment effect which the unadjusted analysis addresses. It is primarily useful in regulatory clinical trials that require analyses to be specified a priori. To illustrate, the method is applied to a study of lung disease with multivariate time-to-event outcomes.

email: b.saville@vanderbilt.edu

A BAYESIAN PHASE I/II DESIGN FOR ONCOLOGY CLINICAL TRIALS OF COMBINATIONAL BIOLOGICAL AGENTS

Chunyan Cai, University of Texas MD Anderson Cancer Center*
Ying Yuan, University of Texas MD Anderson Cancer Center
Yuan Ji, University of Texas MD Anderson Cancer Center

Treating patients with novel biological agents has been a leading trend in oncology. Unlike cytotoxic agents, for which toxicity and efficacy monotonically increase with dose, biological agents may exhibit non-monotonic patterns in their dose-response relationships. To accommodate the patterns of biological agents, we propose a phase I/II trial design to identify the biologically optimal dose combination (BODC), which is defined as the dose combination with the highest efficacy and tolerable toxicity. A change-point model is used to reflect the fact that the dose-toxicity surface of the combinational agents may plateau at higher dose levels, and a flexible logistic model is proposed to accommodate the possible non-monotonic pattern for the dose-efficacy relationship. During the trial, we continuously update the posterior estimates of toxicity and efficacy and assign patients to the most appropriate dose combination. We propose a novel dose-finding algorithm to encourage sufficient exploration of the two-dimensional dose-combination space. Extensive simulation studies show that the proposed design has desirable operating characteristics in identifying the BODC under various patterns of dose-toxicity and dose-efficacy relationships.

email: chunyan.cai@uth.tmc.edu

EMPIRICAL BAYESIAN METHODS FOR ENROLLMENT AND EVENT PROJECTION IN ONCOLOGY TRIALS

Jingyang Zhang, University of Iowa*
Luyan Dai, Boehringer Ingelheim Pharmaceuticals, Inc.
Wei Zhang, Boehringer Ingelheim Pharmaceuticals, Inc.

In clinical trials, the recruitment time is crucial in the design stage, and when the trial outcome is time to event, the design of the trial is mainly based on the number of events. To predict the landmark dates for the interim analyses and final analysis in oncology trials, we summarized the methods from the current literatures and proposed a more general model. The new method incorporates the Poisson-gamma recruitment model into the exponential event prediction model, and both predictions are in the empirical Bayesian setting. The accrual rates in all centers, the event rates and censoring rates in all treatment arms are considered as random variables having a prior gamma distribution. The parameters in prior distributions are estimated through the current data by maximum likelihood method, and the projection for enrollment and event are based on the estimated posterior distributions. Via the demonstration with a real data set together with simulation studies, our new proposed method provides an alternative way to design the clinical trials and update the prediction of the landmark dates for ongoing trials.

email: jingyang-zhang@uiowa.edu

ANALYSIS OF ZERO-INFLATED COUNT DATA FROM CLINICAL TRIALS WITH POTENTIAL DROPOUTS

Jingyuan Yang, Amgen Inc.*
Xiaoming Li, Gilead Sciences, Inc.
Guanghan F. Liu, Merck & Co.

Count of a pre-specified event is an important endpoint for many safety and efficacy clinical trials. The conventional Poisson model might not be ideal due to three potential issues: 1) over-dispersion arising from intra-subject correlation, 2) zero inflation when the pre-specified event is rare, and 3) missing observations due to early dropouts. Negative binomial (NB), Poisson hurdle (PH), and negative binomial hurdle (NBH) models are more appropriate for over-dispersed and/or zero-inflated count data. An offset can be included in these models to adjust for differential exposure duration due to early dropouts. In this paper, we propose new link functions for the hurdle part of a PH/NBH model to facilitate testing for zero-inflation and model selection. The proposed link function particularly improves the model fit of a NBH model when an offset is included to adjust for differential exposure. A simulation study is conducted to compare the existing and proposed models, which are then applied to data from two clinical trials to demonstrate application and interpretation of these methods.

email: jingyuan@amgen.com

A GENERALIZED CONTINUAL REASSESSMENT METHOD FOR TWO-AGENT PHASE I TRIALS

Thomas M. Braun, University of Michigan
Nan Jia, University of Michigan*

We propose a generalized version of the Continual Reassessment Method (CRM), denoted gCRM, for identifying the maximum tolerated combination (MTC) in Phase I trials of two agents. For each dose of one agent, we apply the traditional CRM to study doses of the other agent; each of these CRM designs assumes the same dose-toxicity model, as well as the value of the parameter used in the model. However, each model includes a second parameter that varies among the models in an effort to allow flexibility when modeling the probability of dose-limiting toxicity of all combinations, yet borrow strength among neighboring combinations as well. For example, with a traditional one-parameter logistic model, our approach is seen to lead to a proportional odds logistic model. We incorporate an adaptive Bayesian algorithm to sequentially assign each patient to the most appropriate dose combination, as well as focus patient assignments to a dose combination that has a dose-limiting toxicity (DLT) probability closest to a pre-specified target rate. We test the performance and sensitivity of our method via extensive simulations in various scenarios that are likely to arise in two-agent phase I trials. We also compare the operating characteristics of our approach to several other recently-published approaches.

email: jnan@umich.edu

A HIERARCHICAL BAYESIAN DESIGN IN RANDOMIZED PHASE II CLINICAL TRIALS WITH MULTIPLE SUBGROUPS USING BINARY ENDPOINTS

Qian Shi, Mayo Clinic
Jun Yin, University of Iowa*
Daniel J. Sargent, Mayo Clinic
Charles Erlichman, Mayo Clinic
Rui Qin, Mayo Clinic

Enhanced knowledge of the biologic and genetic basis of disease is re-defining target populations for a new treatment. In oncology, potential target indications for a new therapeutic agent often include various solid tumors and hematologic malignancies that share common signaling pathways. Historically, separate clinical trials of the same agent in each population defined by anatomic site have been conducted in parallel, without any formal mechanism to share information across trials. We proposed a Bayesian hierarchical design to simultaneously test a novel agent in multiple tumors. We assume treatment effects across tumor groups are conditionally exchangeable and correlated while allowing for heterogeneity in control arms. Decision rules are tailored to individual tumor while still maintaining the integrative and sound frequentist study operating characteristics across tumor groups. Posterior probabilities of pre-defined decision rules are estimated via Markov Chain Monte Carlo method. An R package is developed to implement BRugs code and streamline sample size search. Comparing to a separate phase II trial in each tumor groups, the Bayesian hierarchical design reduces sample size by half, therefore, improves efficiency and decreases financial cost.

email: jun-yin@uiowa.edu

VARIABLE SELECTION FOR COVARIATE-ADJUSTED SEMIPARAMETRIC INFERENCE IN RANDOMIZED CLINICAL TRIALS

Shuai Yuan, North Carolina State University*
Helen Zhang, North Carolina State University
Marie Davidian, North Carolina State University

It is well-recognized that a proper covariate-adjusted analysis can improve the efficiency of inference on the treatment effect. However, such covariate adjustment has engendered considerable controversy, as post-hoc selection of covariates may involve subjectivity and lead to biased inference, while prior specification of the adjustment may exclude important variables from consideration. Accordingly, how to select covariates objectively to gain maximal efficiency is of broad interest. We propose and study the use of modern variable selection methods for this purpose in the context of a semiparametric framework, under which variable selection in modeling the relationship between outcome and covariates is separated from estimation of the treatment effect, circumventing the potential for selection bias associated with standard analysis of covariance methods. We demonstrate that such objective variable selection techniques combined with this framework can identify key variables and lead to unbiased and efficient inference on the treatment effect. A critical issue in finite samples is validity of estimators of uncertainty. We propose an approach to estimation of sampling variation of estimated treatment effect and show its superior performance relative to that of existing methods.

email: syuan@ncsu.edu

43. COMPETING RISKS

FRAILTY-BASED COMPETING RISKS MODEL FOR MULTIVARIATE SURVIVAL DATA

Malka Gorfine, Technion – Israel Institute of Technology*
Li Hsu, Fred Hutchinson Cancer Research Center

In this work we provide a new class of frailty-based competing risks models for clustered failure times data. This class is based on expanding the competing risks model of Prentice et al. (1978) to incorporate frailty variates. Nonparametric maximum likelihood estimators (NPMLEs) are proposed. The main advantages of the proposed class of models, in contrast to the existing models, are: (1) the inclusion of covariates; (2) the flexible structure of the dependency among the various types of failure times within a cluster; and (3) the unspecified within-subject dependency structure. The proposed estimation procedure produces the most efficient semiparametric estimators, as opposed to the existing approaches, and it is easy to implement. Simulation studies show that the proposed method performs very well in practical situations.

email: gorfinm@ie.technion.ac.il

SEMIPARAMETRIC ESTIMATION IN THE PROPORTIONAL SUBDISTRIBUTION HAZARDS MODEL WITH MISSING CAUSE OF FAILURE

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

In analyses involving competing risks, the proportional subdistribution hazards regression model of Fine and Gray is commonly used to estimate covariate effects of specific risk factors for disease. In some situations however, the actual cause of failure may be unknown or cannot be determined. To avoid bias, we develop two semiparametric estimators of covariate effects: the inverse probability weighted complete-case estimator and the augmented inverse probability weighted estimator. We study the properties of these estimators analytically and use simulations to compare their small to moderate sample size performance to that of estimators obtained via a multiple imputation method, a naive complete-case analysis, and a method in which missing cases are treated as an extra failure type. We employ the proposed methods to estimate the effects that several risk factors have on the development of coronary heart disease or the occurrence of death related to this disease among individuals infected with the human immunodeficiency virus (HIV).

email: jgy2@pitt.edu

HIERARCHICAL LIKELIHOOD INFERENCE ON CLUSTERED COMPETING RISKS DATA

Nicholas J. Christian, University of Pittsburgh*

Frailties models, an extension of the proportional hazards model, are used to model clustered survival data. In some situations there may be competing risks within a cluster. When this happens the basic frailty model is no longer appropriate. A useful alternative is the cause-specific hazard frailty model. In this work, hierarchical likelihood (h-likelihood) methods are extended to provide a new method for fitting this model. The h-likelihood allows for estimating unobservable heterogeneity, which can be used for modeling the individual effect of treatments in clinical trials. Methods for model selection as well as testing for covariate and clustering effects are also discussed. Simulations demonstrate that the h-likelihood performs well when estimating the cause-specific hazard frailty model assuming a bivariate frailty distribution. A real example from a breast cancer clinical trial is used to demonstrate using h-likelihood inference on clustered competing risks data.

email: njc23@pitt.edu

SUBDISTRIBUTION REGRESSION WITH LEFT-TRUNCATED SEMI-COMPETING RISKS DATA

Ruosha Li, University of Pittsburgh*
Limin Peng, Emory University

Semi-competing risks data frequently arise in biomedical studies when time to a landmark event of disease is subject to dependent censoring by death, the observation of which however is not precluded by the occurrence of the landmark event. In this case, the cumulative incidence function (i.e. subdistribution) for the disease endpoint is often advocated to characterize the disease progression while accounting for the presence of death. In observational studies, left truncation is often present and can greatly complicate the regression analysis based on this cumulative incidence function. In this work, we address this challenge and propose a new semi-parametric regression method, which can flexibly accommodate varying covariate effects and also provide straightforward coefficient interpretation. The proposed methods can be easily implemented via standard statistical software, and the resulting estimators can be shown to have nice asymptotic properties. Our simulation studies show that the proposed methods perform well with realistic sample sizes. An application to the Denmark diabetes registry data demonstrates the practical utility of the proposed methods.

email: rul12@pitt.edu

SIMULATING CLUSTERED COMPETING RISKS DATA

Ruta Brazauskas, Medical College of Wisconsin*
John P. Klein, Medical College of Wisconsin
Jennifer G. Le-Rademacher, Medical College of Wisconsin

We are interested in techniques to simulate competing risks data where individuals within a pair or cluster are associated. Clustered competing risks data arise often in genetic studies, multicenter investigations, and matched-pair studies. Some mechanisms for simulating clustered competing risks data have been considered in the literature. However, most of them produce data where the strength of the dependence between individuals within a cluster is not clear. In this presentation, we will examine techniques to generate bivariate cumulative incidence functions by extending methods used to generate univariate cumulative incidence functions. We will discuss the properties of each technique and provide standard measures of association to assess the degree of dependence in simulated clustered competing risks data.

email: ruta@mcw.edu

ANALYSIS OF DEPENDENTLY CENSORED DATA BASED ON QUANTILE REGRESSION

Shuang Ji, Emory University*
Limin Peng, Emory University
Ruosha Li, University of Pittsburgh
Michael J. Lynn, Emory University

Dependent censoring occurs in many biomedical studies and poses considerable methodological challenges for survival analysis. In this work, we develop a new approach for analyzing dependently censored data by adopting quantile regression models. We formulate covariate effects on the quantiles of the marginal distribution of the event time of interest. Such a modeling strategy can accommodate a more dynamic relationship between covariates and survival time compared to traditional regression models in survival analysis, which usually assume constant covariate effects. We propose estimation and inference procedures, along with an efficient and stable algorithm. We establish the uniform consistency and weak convergence of the resulting estimators. Extensive simulation studies demonstrate good finite-sample performance of the proposed inferential procedures. We illustrate the practical utility of our method via an application to a multicenter clinical trial that compared warfrin and aspirin in treating symptomatic intracranial arterial stenosis.

email: sji@emory.edu

44. FUNCTIONAL DATA ANALYSIS

LONGITUDINAL SURVEY SAMPLING OF FUNCTIONAL DATA

David Degras, DePaul University*

When collections of functional data (i.e. high resolution digitized signals) are too large to be exhaustively observed, survey sampling methods provide an advantageous tradeoff between analysis costs and statistical accuracy. Although it is well known in longitudinal survey theory that replacing the sample over time improves the estimation of population parameters, survey methods for functional data have so far exclusively been based on time-invariant samples. In this work we propose two novel sampling designs for the survey of functional data. These designs produce flexible stratified samples that can be replaced in part or in full at given times. Considering Horvitz-Thompson estimators of the population mean signal, we derive large-sample approximations of the mean and variance of the integrated squared error. We show that frequently replacing the sample dramatically reduces the variance of the estimation error. Further, the periodic reallocation of the sample across the strata can substantially reduce the mean estimation error. In an application to simulated electricity load curves, our sampling designs are seen to compare positively with classical survey methods.

email: ddegrasv@depaul.edu

CORRECTED CONFIDENCE BANDS FOR FUNCTIONAL DATA USING PRINCIPAL COMPONENTS

Jeff Goldsmith, Johns Hopkins Bloomberg School of Public Health*
Sonja Greven, Ludwig-Maximilians-University
Ciprian Crainiceanu, Johns Hopkins Bloomberg School of Public Health

Functional principal components (FPC) analysis is widely used to decompose and express functional observations. Curve estimates implicitly condition on basis functions and other quantities derived from FPC decompositions; however these objects are unknown in practice. In this paper, we propose a method for obtaining correct curve estimates by accounting for uncertainty in FPC decompositions. Additionally, pointwise and simultaneous confidence intervals that account for both model-based and decomposition-based variability are constructed. Standard mixed-model representations of functional expansions are used to construct curve estimates and variances conditional on a specific decomposition. A bootstrap procedure is implemented to understand the distribution of principal components decompositions. Iterated expectation and variance formulas combine both sources of uncertainty by averaging model-based conditional estimates across the distribution of decompositions. Our method compares favorably to competing approaches in simulation studies that include both densely- and sparsely-observed functions. We apply our method to sparse observations of CD4 cell counts and to dense white-matter tract profiles. Code for the analyses and simulations is publicly available.

email: jgoldsmi@jhsph.edu

OPTIMAL SMOOTHING BANDWIDTH SELECTION METHODS FOR FUNCTIONAL DATA

Jingjing Yang, Rice University*
David W. Scott, Rice University
Dennis D. Cox, Rice University

We will develop and test adaptive cross validation and Bayesian methods for selecting optimal smoothing parameters for functional data. Functional data are generally obtained from discrete noisy observations of continuous curves. Typically, smoothing each curve is treated as a single nonparametric regression problem, or a single bandwidth is applied to all curves. We investigate nonparametric regression methods that borrow strength from all of the measured curves, i.e. across all realizations of the functions. An adaptive cross-validation method and a Bayesian method for selecting optimal smoothing parameters will be presented and compared. Simulations and application to real case studies will be presented, along with theoretical justifications for some results. This extended nonparametric regression model could be used widely in many application areas, such as regressing spectrum data, as they are all continuous data curves and could be treated as a function of some variables.

email: jy13@rice.edu

MULTISCALE ADAPTIVE COMPOSITE QUANTILE REGRESSION MODELS FOR NEUROIMAGING DATA

Linglong Kong, University of North Carolina at Chapel Hill*
Hongtu Zhu, University of North Carolina at Chapel Hill

Neuroimaging studies aim to analyze imaging data with complex spatial patterns in a large number of locations (called voxels) on a two-dimensional (2D) surface or in a 3D volume. We propose a multiscale adaptive composite quantile regression model (MACQRM) that has four attractive features: being robustness, being spatial, being hierarchical, and being adaptive. MACQRM utilizes imaging observations from the neighboring voxels of the current voxel and borrows strength from the nearby quantile regressions of the current regression to adaptively calculate parameter estimates and test statistics. Theoretically, we establish consistency and asymptotic normality of the adaptive estimates and the asymptotic distribution of the adaptive test statistics. Our simulation studies and real data analysis confirm that MACQRM significantly outperforms MARM and conventional analyses of imaging data.

email: llkong@bios.unc.edu

ESTIMATION OF FUNCTIONAL CURVE PEAK LOCATIONS FOR DETECTION OF CERVICAL PRE-CANCER

Lu Wang, Rice University*
Dennis D. Cox, Rice University

Cervical cancer is easy to prevent if detected early. We are investigating the use of spectroscopic devices that have been shown to have power to detect cancerous and pre-cancerous lesions. One of them major problems with bio-medical applications of optical spectroscopy has been repeatability of the measurements. Rhodamine is one of the mostly commonly used standards in fluorescence spectroscopy. The measured spectra are functional data with variations due to different devices, measurement conditions, excitation wavelengths and some operational effects. The observed curve peak locations may be shifted by contamination fluorescence. To estimate the true spectral peak locations, we propose a model that incorporates the proportion of light from the eal spectrum, together with the effect of contaminations. Simulation and real data application show that the proposed model provides accurate estimation of intensity peak locations and heights.

email: lw7@rice.edu

LONGITUDINAL FUNCTIONAL REGRESSION MODELS WITH STRUCTURED PENALTIES

Madan G. Kundu, Indiana University School of Medicine*
Jaroslaw Harezlak, Indiana University School of Medicine
Timothy W. Randolph, Fred Hutchinson Cancer Research Center

Functional data are becoming increasingly common and are being gathered longitudinally in many studies. For example, magnetic resonance spectroscopy produces a spectrum which is a mixture of metabolite spectra, instrument noise and baseline profile. Analysis of such data usually proceeds in two disconnected steps: feature extraction and statistical modeling. In contrast, the newly proposed partially empirical eigenvectors for regression (PEER) approach for functional linear models incorporates a priori knowledge via a scientifically-informed penalty operator in the estimation process. We extend the scope of PEER to the longitudinal setting with continuous outcome and a longitudinal functional covariate. The method presented in this paper: 1) takes into account the external information and 2) allows time-varying parameter function. At each time-point the parameter function is decomposed into several time-invariant components with the time dependence entering through their coefficients. We derive the precision and accuracy of the estimates and discuss their connection with the generalized singular value decomposition. A real data and simulations are used to illustrate the concepts.

email: mgkundu@iupui.edu

FUNCTIONAL MIXED-EFFECTS MODELS FOR MULTIPLE OUTCOMES

Stephanie A. Kliethermes, University of Iowa*
Jacob J. Oleson, University of Iowa

Current research in functional data analysis (FDA) methodology involves analyzing longitudinal data where the primary unit is a curve over time. FDA assumes the data are measurements of smooth, infinite-dimensional curves and observations at each time point are noise along the underlying curve. Yet, individuals often have limited observations representing their true curves. Functional mixed effects (FME) models were proposed to model these curves and methods have been developed to handle sparse observations. However, the problem of modeling curves in situations where subjects have multiple outcomes over time, resulting in correlated curves, remains. Our research accounts for this correlation by using Bayesian techniques to expand existing FME models. We focus on a study at The University of Iowa where the hearing abilities of individuals with new cochlear implants were tested regularly for two years post-implantation. Each subject was tested under various conditions (hearing ability in the presence/absence of hearing aids) resulting in multiple outcomes per visit. We determine the underlying growth (in hearing ability) curves for each condition. Our multiple outcome FME model estimates these curves while accounting for correlation between outcomes and thus helps to determine optimal hearing conditions for implanted individuals.

email: stephanie-kliethermes@uiowa.edu

45. GENOME-WIDE ASSOCIATION STUDIES

GENOME-WIDE ASSOCIATION ANALYSIS FOR MULTIPLE CONTINUOUS SECONDARY PHENOTYPES

Elizabeth D. Schifano, Harvard School of Public Health*
Lin Li, 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. Multiple phenotypes often measure the same underlying trait. By taking advantage of correlation across outcomes, one could potentially gain statistical power. Since continuous phenotypes are likely to be measured on different scales, a scaled marginal model for testing and estimating the shared SNP effect on the multiple phenotypes is proposed to borrow additional strength across outcomes. We extend the scaled marginal model for testing and estimating this shared SNP effect on multiple secondary phenotypes in case-control studies by using weighted estimating equations. This approach does not require correct specification of the within-subject correlation, and simultaneously accounts for case-control ascertainment. We perform simulation studies to show that our one-degree-of-freedom test for shared SNP effect on multiple related outcomes is more powerful than either testing the outcomes separately or testing the outcomes jointly with a traditional multiple-degree-of-freedom test. The proposed method is applied to a case-control lung cancer GWAS to investigate SNP associations with multiple secondary phenotypes related to smoking behavior.

email: eschifan@hsph.harvard.edu

LONGITUDINAL GENETIC ANALYSIS OF QUANTITATIVE TRAITS

Ruzong Fan, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*

Longitudinal genetic studies provide a very valuable resource for exploring key genetic and environmental factors that affect complex traits over time. Genetic analysis of longitudinal genetic data that incorporates temporal variations is important for understanding genetic architecture and biological variations of common complex diseases. It may provide a powerful tool for identifying genetic determinants of complex diseases, and for understanding at which stage of human development genetic determinants are important. Moreover, important environmental factors which are associated with complex diseases, such as diet, socio-economic status, and smoking status, can be identified. Although they are important, there are limited statistical models and methods to analyze longitudinal human genetic data. For instance, there is no combined linkage and association analysis of the Framingham Heart Study data. The reason is that there are no longitudinal statistical models and software for a joint linkage and association study of temporal quantitative traits of complex diseases. In this project, we develop a general framework for longitudinal analysis of function-valued quantitative traits for analyzing genetic family data. Simulation studies are performed to evaluate the performance of the models and methods.

email: fanr@mail.nih.gov

INCORPORATING GROUP CORRELATIONS IN GENOME-WIDE ASSOCIATION STUDIES USING SMOOTHED GROUP LASSO

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

In genome-wide association studies, penalization is becoming an important approach for identifying genetic markers associated with disease. Motivated by the fact that there exists natural grouping structure in SNPs and more importantly such groups are correlated, we propose a new penalization method for group

variable selection which can properly accommodate the correlation between adjacent groups. This method is based on a combination of the group Lasso penalty and a quadratic penalty on difference of regression coefficients of adjacent groups. The new method is referred to as Smoothed Group Lasso, or SGL. It encourages group sparsity and smoothes regression coefficients for adjacent groups. We derive a group coordinate descent algorithm for computing the solution path. The SGL method is further extended to logistic regression for binary response. With the assistance of MM algorithm, the logistic regression model with SGL penalty turns out to be an iteratively penalized least-square problem. Principal components are used to reduce dimensionality locally within groups. Simulation studies are used to evaluate the finite sample performance. Comparison with group Lasso shows that SGL is more effective in selecting true groups. We also analyze a rheumatoid arthritis data by applying the SGL method under logistic regression model.

email: jin.liu.jl2329@yale.edu

A PENALIZED LIKELIHOOD APPROACH FOR PHARMACOGENETIC STUDIES VIA UNDERSTANDING HAPLOTYPE EFFECT STRUCTURES FOR GENE AND GENE-DRUG INTERACTIONS

Megan L. Neely, Duke University*
Howard D. Bondell, North Carolina State University
Jung-Ying Tzeng, North Carolina State University

Pharmacogenetics investigates the relationship between genetic variation and the variation in individuals' response to drugs. Often, gene-drug interactions play a primary role in this response, and identifying these effects aid in the development of individualized treatment regimes. Haplotypes can hold key information in understanding the association between genetic variation and drug response. However, the standard approach for haplotype-based analyses does not directly address the research questions dictated by individualized medicine. A complimentary post-hoc analysis is required, and this analysis is usually under powered after adjusting for multiplicity and may lead to contradictory conclusions. In this work, we propose a penalized likelihood approach that is able to overcome the drawbacks of the standard approach and yield the desired personalized output. We demonstrate the utility of our method by applying it to the Scottish Randomized Trial in Ovarian Cancer. We also conducted simulation studies and showed that the proposed penalized method has comparable or more power than the standard approach and maintains low Type I error rates for both binary and quantitative drug responses. The largest power gains are seen when haplotype frequency is low, the difference in effect size is small, or the true relationship among the drugs is more complex.

email: megan.koehler@duke.edu

THE EFFECT OF POPULATION STRATIFICATION ON ASSOCIATION STUDIES WITH NEXT GENERATION SEQUENCING

Qianying Liu, University of Chicago*
Lin Chen, University of Chicago
Dan L. Nicolae, University of Chicago

Population stratification can lead to spurious association between variants and disease status in a case-control study. In the past decade, great efforts have been extended to detect and to adjust for population stratification, leading to efficient solutions for genome-wide association studies. Here we examine, both theoretically and empirically, the effect of population stratification on gene-based association tests in the context of next generation sequencing studies. With extensive simulations, we evaluate the performance of two commonly used approaches for adjusting population stratification -- genomic control and principal component analysis and show how commonly used strategies for single-SNP tests do not work properly for pooled analyses of rare variants.

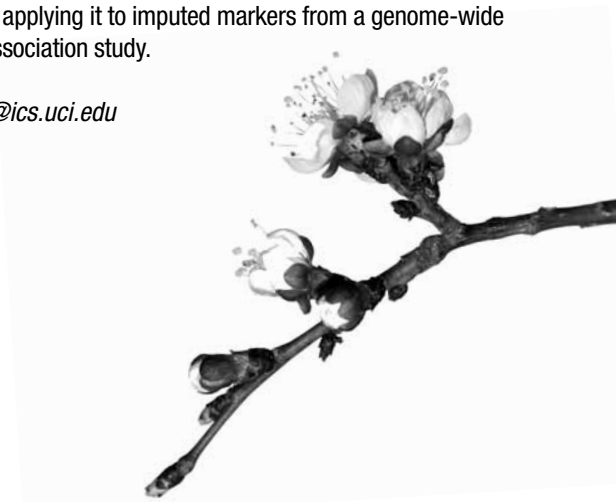
email: qianyingliu05@gmail.com

FAMILY-BASED ASSOCIATION TESTS USING GENOTYPE DATA WITH UNCERTAINTY

Zhaoxia Yu, University of California, Irvine*

Family-based association studies have been widely used to identify association between diseases and genetic markers. It is known that genotyping uncertainty is inherent in both directly genotyped or sequenced DNA variations and imputed data in silico. The uncertainty can negatively impact both the power and Type I error rates of family-based association studies. Compared to studies using unrelated subjects, there are very few methods that address the issue of genotyping uncertainty for family-based designs. Here we propose a new test to address the challenges in using uncertainty genotype data in family based association studies. Our simulations show that compared to the conventional strategy and an alternative test, our new test has an improved performance in the presence of substantial uncertainty, and has a similar performance when the uncertainty level is low. We also demonstrate the advantages of our new method by applying it to imputed markers from a genome-wide case-parents association study.

email: zhaoxia@ics.uci.edu



■ = Student Award Winner | ✶ = Presenter

46. STATISTICAL MODELS FOR OMICS DATA

DIFFERENTIAL PRINCIPAL COMPONENT ANALYSIS OF ChIP-Seq

Hongkai Ji, Johns Hopkins University Bloomberg School of Public Health*
Yang Ning, Johns Hopkins University Bloomberg School of Public Health

We propose Differential Principal Component Analysis (dPCA) for characterizing differences between two biological conditions with respect to multiple ChIP-seq data sets. dPCA describes major differential patterns between two conditions using a small number of principal components. Each component corresponds to a multi-dataset covariation pattern shared by many genomic loci. The analysis prioritizes genomic loci based on each pattern, and for each pattern, it identifies loci with significant between-condition changes after considering variability among replicate samples. This approach provides an integrated solution to dimension reduction, unsupervised pattern discovery, and statistical inference. We demonstrate dPCA through analyses of differential chromatin patterns at transcription factor binding sites and human promoters using ENCODE data.

email: hji@jhsph.edu

BAYESIAN HIERARCHICAL FUNCTIONAL MODELS FOR HIGH-DIMENSIONAL GENOMICS DATA

Veera Baladandayuthapani, University of Texas MD Anderson Cancer Center*
Jeffrey S. Morris, University of Texas MD Anderson Cancer Center
Yuan Ji, University of Texas MD Anderson Cancer Center

Recent advances in genomic profiling techniques such as array-based comparative genomic hybridization (array-CGH) and SNP arrays provide a high-throughput, high-resolution method to measure relative changes in DNA copy number. These experiments typically yield data consisting of profiles of copy number changes of hundreds/thousands of markers across the whole chromosomal map. Modeling and inference in such studies is challenging not only due to high-dimensionality but also due to presence of serial correlation of the markers along the genome. Using genome continuum models as a general principle we present a class of Bayesian methods to model these genomic profiles using functional data analysis approaches. Our methods allow for simultaneous characterization of these high-dimensional functions, borrowing strength between replicated functions and detection of local features in the data to answer several important biological questions using such data. The methods are illustrated using several real and simulated datasets.

email: veera@mdanderson.org

A BAYESIAN GRAPHICAL MODEL FOR ChIP-Seq DATA ON HISTONE MODIFICATIONS

Peter Mueller, University of Texas at Austin*
Riten Mitra, University of Texas MD Anderson Cancer Center
Shoudan Liang, University of Texas MD Anderson Cancer Center
Lu Yue, University of Texas MD Anderson Cancer Center
Yuan Ji, University of Texas MD Anderson Cancer Center

Histone modifications (HMs) are an important post-translational feature. Different types of HMs are believed to co-regulate biological processes such as gene expression, and therefore are intrinsically dependent on each other. We develop inference for this complex biological network of HMs based on a graphical model for the dependence structure across HMs. A critical computational hurdle in the inference for the proposed graphical model is the evaluation of a normalization constant in an autologistic model that builds on the graphical model. We tackle the problem by Monte Carlo evaluation of ratios of normalization constants. We carry out a set of simulations to validate the proposed approach and to compare it with a standard approach using Bayesian networks. We report inference on HM dependence in a case study with ChIP-Seq data from a next-generation sequencing experiment. An important feature of our approach is that we can report coherent probabilities and estimates related to any event or parameter of interest, including honest uncertainties. Posterior inference is obtained from a joint probability model on latent indicators for the recorded HMs. We illustrate this in the motivating case study. An R package including an implementation of posterior simulation in C is available.

email: pmueller@math.utexas.edu

A BAYESIAN NETWORK ANALYSIS FOR SINGLE-CELL MASS CYTOMETRY DATA

Riten Mitra, University of Texas MD Anderson Cancer Center
Yuan Ji, University of Texas MD Anderson Cancer Center*
Peter Mueller, University of Texas at Austin

We present a statistical inference for protein network based on single- cell mass cytometry data (Bendall et al., 2011, Science). The single-cell data allow investigators to study cellular protein activities at the cell level as opposed to the sample level, thus potentially eliminating biases caused by biological samples containing mixed types of cells. The goal of our analysis is to infer the network dependence structures of selected cellular protein markers for different types cells. We also report differential network analysis comparing 1) pairs of cell types and 2) the same cell type before and after various drug treatments. The statistical model is based on a general Bayesian graphic inference with covariates. Properties and limitations of the proposed models will be discussed.

email: koaeraser@gmail.com

47. TWEEDIE AWARD

STATISTICAL LEARNING WITH HIGH-DIMENSIONAL DATA

Hui Zou, University of Minnesota*

High-dimensionality has revolutionized the landscape of statistical inference and learning. After a brief literature review I will use two examples to illustrate some strategies to exploit the sparsity assumption in high-dimensional learning. In the first example I will discuss the problem of sparse discriminant analysis which has received a lot of attention in the past decade. This is a classical supervised learning problem. Some fundamental drawbacks of existing proposal will be pointed out and a new approach to sparse discriminant analysis will be presented and demonstrated by theoretical analysis and many numerical examples. The second example concerns learning graphical models with non-Gaussian data. Under normality assumption graphical model learning is often formulated as estimating the precision matrix of a multivariate normal distribution. However, the observed data are often skewed or have heavy tails. To deal with the non-normality issue I will introduce a much more flexible graphical model and new estimation methods will be presented together with theoretical and numerical results.

email: zouxx019@umn.edu

ADAPTIVE ESTIMATION OF LARGE COVARIANCE MATRICES

Tony Cai, University of Pennsylvania*
Ming Yuan, Georgia Institute of Technology

Estimation of large covariance matrices has drawn considerable recent attention and the theoretical focus so far is mainly on developing a minimax theory over a fixed parameter space. In this paper, we consider adaptive covariance matrix estimation where the goal is to construct a single procedure which is minimax rate optimal simultaneously over each parameter space in a large collection. A fully data-driven block thresholding estimator is proposed. The estimator is constructed by carefully dividing the sample covariance matrix into blocks and then simultaneously estimating the entries in a block by thresholding. The estimator is shown to be optimally rate adaptive over a wide range of bandable covariance matrices.

email: tcai@wharton.upenn.edu

48. RECENT DEVELOPMENT IN OPTIMAL TREATMENT STRATEGIES — ESTIMATION, SELECTION, AND INFERENCE

EVALUATING OPTIMAL TREATMENT POLICIES BASED ON GENE EXPRESSION PROFILES

Ian McKeague, Columbia University*
Min Qian, Columbia University

This talk discusses optimal treatment policies based on interactions between treatment and gene expression, and determined by thresholds in gene expression at a small number of loci along a chromosome. Such treatment policies are easier to interpret and implement (in bioassays, say) than policies based on complete gene expression profiles. By formulating the statistical problem in terms of a sparse functional linear regression model, we show how data from randomized clinical trails can be used to simultaneously evaluate the effectiveness of the treatment policies (measured in terms of mean outcome when all patients follow the policy), and to locate genes that optimize the interaction effect over competing treatments.

email: im2131@columbia.edu

ITERATIVE OUTCOME WEIGHTED LEARNING FOR ESTIMATING OPTIMAL DYNAMIC TREATMENT REGIME

Donglin Zeng, University of North Carolina at Chapel Hill*
Yingqi Zhao, University of North Carolina at Chapel Hill
Michael Kosorok, University of North Carolina at Chapel Hill

Traditional approaches to select dynamic treatment regime in sequentially multiple randomization trials include Q-learning and A-learning. These approaches require modeling the relationship between rewards or regrets, treatments and pretreatment prognostic variables, and then inverting the model to obtain the optimal treatment assignments. However, in the presence of large number of predictors, they are likely to overfit the model and lead to a suboptimal treatment policy. In this work, we propose a novel approach by directly maximizing the expected rewards associated with each treatment policy. For single-decision setup, we transform the maximization into an outcome-weighted learning framework and apply a weighted support vector machine for estimation. While with multi-decision setup, we utilize an iterative outcome weighted learning technique for estimation. The proposed procedure is shown to lean to optimal treatment decision rule and can be easily implemented with existing software. Finally, we illustrate the approach via numerical studies.

email: dzeng@email.unc.edu

UP-FRONT VS. SEQUENTIAL RANDOMIZATIONS FOR INFERENCE ON ADAPTIVE TREATMENT STRATEGIES

Abdus S. Wahed, University of Pittsburgh*
Jin.H.Ko, University of Pittsburgh

Dynamic treatment regimes aka adaptive treatment strategies are useful in the treatment of chronic diseases such as AIDS and cancer, since they allow tailoring the treatment to patient's need and disease status. We consider two randomization schemes for clinical trials that are commonly used to design studies comparing adaptive treatment strategies, namely, up-front randomization and sequential randomization. Up-front randomization is the classical method of randomization where patients are randomized at the beginning of the study to pre-specified treatment strategies. In sequentially randomized trials, patients are randomized sequentially to available treatment options over the duration of the therapy as they become eligible to receive subsequent treatments. We compare the efficiency and the power of the traditional up-front randomized trials to that of sequentially randomized trials designed for comparing adaptive treatment strategies based on a continuous outcome. The analytical and simulation results indicate that, when properly analyzed, sequentially randomized trials are more efficient and powerful than up-front randomized trials.

email: waheda@edc.pitt.edu

INFERENCE FOR DYNAMIC TREATMENT REGIMES

Eric B. Laber, North Carolina State University*
Daniel J. Lizotte, University of Waterloo
Min Qian, Columbia University
Susan A. Murphy, University of Michigan

Dynamic treatment regimes are increasingly being used to operationalize sequential clinical decision making associated with patient care. Common approaches to constructing a dynamic treatment regime from data, such as Q-learning, employ non-smooth functionals of the data. Therefore, simple inferential tasks such as constructing a confidence interval for the parameters in the Q-function are complicated by nonregular asymptotics under certain commonly-encountered generative models. Methods that ignore this nonregularity can suffer from poor performance in small samples. We construct confidence intervals for the parameters in the Q-function by first constructing smooth, data-dependent, upper and lower bounds on these parameters and then applying the bootstrap. The confidence interval is adaptive in that it is conservative for nonregular generative models, but achieves asymptotically exact coverage elsewhere. The small sample performance of the method is evaluated on a series of examples and compares favorably to previously published competitors. Finally, we illustrate the method using data from the Adaptive Interventions for Children with ADHD study (Pelham et al. 2008).

email: laber@stat.ncsu.edu

49. CHALLENGING ISSUES IN FUNCTIONAL CONNECTIVITY ANALYSIS

PERSISTENT HOMOLOGICAL NETWORK MODELING VIA GRAPH FILTRATION

Moo K. Chung, University of Wisconsin-Madison*

Brain connectivity has been usually modeled as a network graph. The whole brain region can be parcellated into disjoint regions, which serve as the nodes of the network. Functional imaging such as fMRI and PET provides additional information of how one region is connected to another via a connectivity matrix. The connectivity matrix is then thresholded to produce a binarized adjacency matrix, which is further used in constructing a graph. The main problem with the standard framework is the arbitrariness of thresholding connectivity matrix. The topological parameters such as sparsity and clustering coefficients change substantially depending on the level of threshold. The problems of arbitrary thresholding can be avoided if we do not use any thresholding in building the network. So the question is whether if it is possible to construct and model a network graph without any thresholding. In this talk, we present a novel network graph modeling technique motivated by persistent homology, which looks at the persistent topological features of changing networks when the threshold changes. We have applied the method in characterizing abnormal PET and DTI connectivity in autism.

email: mkchung@wisc.edu

PREDICTING NEUROLOGICAL DISORDERS USING FUNCTIONAL AND STRUCTURAL BRAIN IMAGING DATA

Brian S. Caffo, Johns Hopkins University*
Ciprian Crainiceanu, Johns Hopkins University
Han Liu, Johns Hopkins University
Ani Eloyan, Johns Hopkins University
John Muschelli, Johns Hopkins University
Fang Han, Johns Hopkins University
Tuo Zhao, Johns Hopkins University

In this talk we overview methodology for predicting clinical outcomes, and especially neurological disorders, using functional and structural brain imaging data. We focus on resting state functional connectivity data via fMRI as well as structural imaging data via T1 MRI and diffusion weighted MRI. We consider these modalities and variety of methods for feature extraction and prediction. We apply the methodology to developmental disorders, particularly attention deficit hyperactivity, cognitive impairment and Alzheimer's disease and multiple sclerosis.

email: bcaffo@jhsph.edu

FUNCTIONAL CONNECTIVITY THROUGH COLOR INDEPENDENT COMPONENT ANALYSIS

Haipeng Shen, University of North Carolina at Chapel Hill*

Independent component analysis (ICA) is an effective data-driven method for blind source separation. It has been successfully applied to separate source signals of interest from their mixtures. Most existing ICA procedures are carried out by relying solely on the estimation of the marginal density functions. In many applications, correlation structures within each source also play an important role besides the marginal distributions. One important example is functional magnetic resonance imaging (fMRI) analysis where the brain-function-related signals are temporally correlated. I shall talk about a novel ICA approach that fully exploits the correlation structures within the source signals. Specifically, we propose to estimate the spectral density functions of the source signals instead of their marginal density functions. Our methodology is described and implemented using spectral density functions from frequently used time series models. The time series parameters and the mixing matrix are estimated via maximizing the Whittle likelihood function. The performance of the proposed method will be illustrated through simulation studies and a real fMRI application. The numerical results indicate that our approach outperforms several popular methods.

email: haipeng@email.unc.edu

SPATIAL AND ADAPTIVE MODELS FOR BRAIN FUNCTIONAL CONNECTIVITY

Hongtu Zhu, University of North Carolina at Chapel Hill*
Japing Wang, Princeton University

In this talk, we develop several statistical models for characterize brain functional connectivity among different regions of interest at the population studies. We consider modeling in both time domain and frequency domain and develop effective methods to account for both spatial and temporal structure and heterogeneity across subjects. We pose potential solutions as well as future challenges. We apply the methods to studies of Alzheimer's disease.

email: hzhu@bios.unc.edu

50. RECENT DEVELOPMENTS IN SUBGROUP ANALYSIS IN RANDOMIZED CLINICAL TRIALS

KEY STATISTICAL CONSIDERATIONS FOR CLINICAL TRIALS WITH TAILORING OBJECTIVES

Alex Dmitrienko, Quintiles*
Brian Millen, Eli Lilly and Company

This talk focuses on clinical trials pursuing tailoring objectives, eg, include evaluation of treatment effects in focused subpopulations (defined by demographics, clinical or genetic markers) in addition to standard analyses in the overall population. Inferences in the subpopulations are independent of inferences in the overall population and thus may result in regulatory claims even if there is no evidence of a beneficial effect in the overall population. We provide a summary of statistical methods used in tailored therapy trials, including methods for Type I error rate control and analysis considerations to support labeling.

email: alex.dmitrienko@quintiles.com

PREDICTIVE ANALYSIS OF CLINICAL TRIALS

Richard M. Simon, National Cancer Institute, National Institutes of Health*

Developments in genomics and tumor biology have clearly indicated that many if not most conventional cancer diagnoses are heterogeneous with regard to their causative mutations and their response to therapy. The standard paradigm of broad eligibility randomized clinical trials with conclusions based primarily on overall average treatment effect has a less compelling scientific basis than previously and is not appropriate with the new generation of molecularly targeted treatments. The usual approach to subset analysis, however, is also problematic and does not provide a reliable basis for personalized predictive medicine. In this talk I will describe and illustrate a new approach to the analysis of clinical trials which provides an internally validated predictive classifier that can be used as a decision to for selecting treatments for individual patients based on covariate values. The method is based on the Cross-validated adaptive signature design developed by Freidlin, Jiang and Simon (Clinical Cancer Research 16:691, 2010) but can be used broadly and is not restricted to the oncology or gene expression profiling context in which it was originally reported.

email: rsimon@nih.gov

■ = Student Award Winner | * = Presenter

MULTIPLICITY CONSIDERATIONS FOR HYPOTHESES TESTING FOR A TARGETED SUBGROUP TRIAL DESIGN

Mohammad F. Huque, U.S. Food and Drug Administration*
Mohammed Alosh, U.S. Food and Drug Administration

Clinical trials usually assess average treatment effect for the study populations of the trial. However, it is well known that for some clinical indications a targeted subgroup of patients of the trial may benefit more from the treatment than the rest of the patients of the trial. In that case, it is of benefit to design a trial with the goal of establishing efficacy claims for the total patient population of the trial as well as for the targeted subgroup. In this presentation, we address multiplicity considerations for hypotheses testing for such a targeted subgroup trial design. The method of testing to be presented also takes into account the efficacy results for the complimentary subgroup of patients who are not in the targeted subgroup.

email: mohammad.huque@fda.hhs.gov

51. RECENT ADVANCES IN METHODOLOGY FOR THE ANALYSIS OF FAILURE TIME DATA

MARGINAL ADDITIVE HAZARDS MODEL FOR CASE-COHORT STUDIES WITH MULTIPLE DISEASE OUTCOMES

Sangwook Kang, University of Connecticut
Jianwen Cai, University of North Carolina at Chapel Hill*
Lloyd Chambless, University of North Carolina at Chapel Hill

We consider fitting marginal additive hazards regression models for case-cohort studies with multiple disease outcomes. Most modern analyses of survival data focus on multiplicative models for relative risk using proportional hazards models. However, in many biomedical studies, the proportional hazards assumption might not hold or the investigators are often interested in risk differences. The additive hazards model, which model the risk differences, has often been suggested as an alternative to the proportional hazards model. We consider a weighted estimating equation approach for the estimation of model parameters. The asymptotic properties of the proposed estimators are derived and their finite sample properties are assessed via simulation studies. The proposed method is applied to the Atherosclerosis Risk in Communities (ARIC) Study for illustration.

email: cai@bios.unc.edu

STATISTICAL METHODS FOR ASSESSING URGENCY AND TRANSPLANT BENEFIT IN THE PRESENCE OF DEPENDENT CENSORING

Susan Murray, University of Michigan*
Fang Xiang, University of Michigan

Lung allocation priority has shifted from a first-come first-served basis to an algorithm based on estimated urgency and lung transplant benefit. As part of this change, survival outcomes for candidates are now dependently censored according to a daily changing lung allocation score (LAS). We present methods for analysis of transplant benefit and urgency assessment in this new era of transplantation data, where dependent censoring must be accounted for. Multivariate analysis of restricted means for one-year urgency and benefit based on an adjusted pseudo observation approach and an adjusted multiple imputation approach are given and compared. The multiple imputation approach incorporates observed residuals from a restricted mean model fit to the data. This latter approach demonstrates greater precision in estimation and the added advantage of being able to quickly produce adjusted survival estimates, restricted means, two-sample tests and other analyses of interest. When applied to the current lung candidate data, both methods dramatically change estimates of allocation scores when compared to methods that do not account for dependent censoring. Historically known high urgency profiles are better identified using the newer methods. Finally, some thoughts on estimation of transplant urgency in the presence of long-term follow-up data are given.

email: skmurray@umich.edu

SEMIPARAMETRICALLY EFFICIENT TREATMENT EFFECT ESTIMATION IN THE ANALYSIS OF RECURRENT EVENTS

Adin-Cristian Andrei, Northwestern University*

In randomized clinical trials, participants may experience a succession of landmark events. Examples include pulmonary exacerbation episodes, post-treatment (re)hospitalizations or disease reoccurrences. The joint analysis of the inter-event times is an important step towards better understanding the disease process and/or the treatment effect. Oftentimes, large numbers of covariates are also recorded, including patient demographics, medical history, and other baseline and follow-up measures. Improved inference efficiency for the inter-event times distribution could be achieved by incorporating such covariate information. We achieve this by using the semiparametric efficiency theory and develop a large class of tests that allow for more efficient comparisons of the inter-event times distributions between treatment groups. In addition, extensions of this methodology to observational studies are discussed. Simulation studies and applications demonstrate the practical advantages of this approach.

email: aandrei@nmh.org

ESTIMATING TREATMENT EFFECTS FROM A RANDOMIZED CLINICAL TRIAL IN THE PRESENCE OF POST-STUDY TREATMENT

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

In randomized clinical trials involving survival time, a challenge that arises frequently, for example in cancer studies (Manegold et al, 2005), is that during follow up subjects may initiate post-study treatment (PST) after discontinuing from study treatment. Whether and when to start PST may depend on time-dependent confounders and their existence makes inferences from usual approaches, e.g., intent-to-treat analysis or analysis including PST as a time-dependent covariate, lose causal interpretation. Marginal structural Cox's model and method based on inverse probability of treatment weighting (IPTW) has been proposed to account for PST in the presence of time-dependent confounders. IPTW method tends to yield estimators that are of large variance and not stable. In this paper, we adopt the marginal structural Cox's model and propose a method that improves the usual IPTW method. The proposed method improves efficiency by taking full advantage of the study design and exploiting the fact that the study treatment is independent of baseline covariates, guaranteed by randomization. The proposed estimator is consistent and asymptotically normal when the model for PST is correctly specified. The finite-sample performance of the proposed method is demonstrated via simulation studies and by application to data from a cancer clinical trial.

email: mzhangst@umich.edu

52. NEW METHODS AND THEORY IN FUNCTIONAL/LONGITUDINAL DATA ANALYSIS

SPLINE CONFIDENCE BANDS FOR FUNCTIONAL DERIVATIVES

Guanqun Cao, Michigan State University*
Jing Wang, University of Illinois at Chicago
Li Wang, University of Georgia
David Todem, Michigan State University

We develop in this paper a new procedure to construct simultaneous confidence bands for derivatives of mean curves in functional data analysis. The technique involves polynomial splines that provide an approximation to the derivatives of the mean functions, the covariance functions and the associated eigenfunctions. We show that the proposed procedure has desirable statistical properties. In particular, we first show that the proposed estimators of derivatives of the mean curves are semiparametrically efficient. Second, we establish consistency results for derivatives of covariance functions and their eigenfunctions. Most importantly, we show that the proposed spline confidence bands are asymptotically efficient as if all random trajectories were observed with no error. Finally, the confidence band procedure is illustrated through numerical simulation studies and a real life example.

email: cao@stt.msu.edu

GENERALIZED FUNCTIONAL LINEAR REGRESSION

Xiao Wang, Purdue University
Pang Du, Virginia Tech*

In this paper, we consider generalized regression model where responses come from a distribution of exponential family and the predictor is a functional variable. The coefficient function is estimated through a smoothness regularization approach. Under weaker conditions than those for the function principal component approach, we obtain the minimax rates of convergence and show that smoothness regularized estimators achieve the optimal rates of convergence for both prediction and estimation. Our simulations and application examples demonstrate the use of the proposed generalized functional regression method.

email: pangdu@vt.edu

REGULARIZED SMOOTHING IN FUNCTIONAL LINEAR MODELS

Toshiya Hoshikawa, University of Michigan*
Tailen Hsing, University of Michigan

In this talk we consider nonparametric estimation of regression slope functions in functional linear regression models. Similar models have received a considerable attention in the literature. We consider various cases according to the nature of the predictor and data scheme. The estimators will be derived in the appropriate spaces of functions by penalized least squares. Asymptotic theories for the procedures will be provided.

email: toshiyah@umich.edu

SIMULTANEOUS VARIABLE SELECTION AND ESTIMATION IN SEMIPARAMETRIC MODELING OF LONGITUDINAL / CLUSTERED DATA

Shujie Ma, University of California-Riverside
Qiongxia Song, University of Texas at Dallas
Lily Wang, University of Georgia*

We consider the problem of simultaneous variable selection and estimation in additive partially linear models for longitudinal/ clustered data. We propose an estimation procedure via polynomial splines to estimate the nonparametric components and apply

proper penalty functions to achieve sparsity in the linear part. Under reasonable conditions, we obtain the asymptotic normality of the estimators for the linear components and the consistency of the estimators for the nonparametric components. We further demonstrate that, with proper choice of the regularization parameter, the penalized estimators of the nonzero coefficients achieve the asymptotic oracle property. The finite sample behavior of the penalized estimators is evaluated with simulation studies and illustrated by a longitudinal CD4 cell count dataset.

email: lilywang@uga.edu

ROBUST REGULARIZED SINGULAR VALUE DECOMPOSITION FOR TWO WAY FUNCTIONAL DATA

Lingsong Zhang, Purdue University*
Haipeng Shen, University of North Carolina at Chapel Hill
Jianhua Huang, Texas A&M University

We develop a robust regularized singular value decomposition method for analyzing two-way functional data. The research is primarily motivated by the important application of modeling human mortality as a smooth two-way function of age group and year. Our method naturally combines two-way roughness penalization and robust regression. The amount of smoothing regularization can be adaptively selected using generalized cross-validation. The advantages of the developed method are shown via the mortality rate modeling application and extensive simulation studies.

email: lingsong@purdue.edu

53. MULTIVARIATE METHODS IN HIGH DIMENSIONAL DATA

A CALIBRATED MULTICLASS EXTENSION OF ADABOOST

Daniel B. Rubin, U.S. Food and Drug Administration*

We propose a new extension of AdaBoost for classification problems with more than two classes. The method generalizes the statistical view of boosting by fitting a weak learner to iteratively reweighted data to perform forward stagewise modeling, using a multiclass exponential loss function as a surrogate for the nonconvex misclassification loss. The algorithm differs from previous extensions of AdaBoost with multiclass weak learners in that even without placing restrictions on conditional class probabilities the surrogate loss function meets the classification calibration condition, so convergence to the optimal surrogate risk guarantees convergence to the optimal misclassification risk. Numerical experiments show the technique leads to good performance on benchmark problems.

email: daniel.rubin@fda.hhs.gov

PREDICTING MORTALITY IN AN ELDERLY POPULATION USING MACHINE LEARNING

Sherri Rose, Johns Hopkins Bloomberg School of Public Health*

Standard practice for mortality prediction often relies on parametric regression methods. We use the machine learning algorithm super learner in the National Institute of Aging funded Study of Physical Performance and Age-Related Changes in Sonomans (SPPARCS) to predict death among 2066 residents of Sonoma, CA aged 54 years and over. Covariates in the SPPARCS data include self-rated health status, leaisure-time physical activity score, history of cardiac events, and chronic health conditions at baseline, among others. The super learner we implement is a flexible machine learning approach that combines multiple algorithms into a single algorithm, and returns a prediction function with the best cross-validated mean squared error.

email: srose@jhsph.edu

EFFICIENT MULTI-MARKER TESTS FOR ASSOCIATION IN CASE-CONTROL STUDIES

Margaret A. Taub, Johns Hopkins University*
Holger Schwender, TU Dortmund University, Dortmund, Germany
Ingo Ruczinski, Johns Hopkins University
Thomas A. Louis, Johns Hopkins University

As case-control studies employ increasingly dense panels of genetic markers, the linkage disequilibrium (LD) between nearby markers can induce correlation between test statistics for tests of association between genotype and case-control status. Intuitively, a more powerful test than an individual marker test should be possible by taking advantage of this correlation structure. Here, we discuss methods for forming a region-based test statistic, either through an optimal linear combination of score test z-statistics or by taking the maximum of the z-statistics over a window of interest. We present results illustrating the performance of these methods under different models of LD and minor allele frequency distributions.

email: mtaub@jhsph.edu

ESTIMATION OF A NON-PARAMETRIC VARIABLE IMPORTANCE MEASURE OF A CONTINUOUS EXPOSURE

Antoine Chambaz, Université Paris Descartes and CNRS*
Pierre Neuvial, Université d'Evry Val d'Essonne
Mark J. van der Laan, University of California, Berkeley

In this talk, we will define a new class of statistical parameters which we call non-parametric variable importance (NPVI) measures. They extend the notion of variable importance measure of a discrete "cause" onto an "effect" accounting for potential confounders to the case where the "cause" is continuous. They are non-parametric in the sense that it is not necessary to assume that a specific semi-parametric model holds. We will show *how* to carry out the estimation of such a NPVI measure following the targeted minimum loss estimation (TMLE) methodology. Some important asymptotic properties of the TMLE estimator (robustness, asymptotic normality) will be stated. The talk will be illustrated with a simulation study inspired by biological question of interest and a dataset from the Cancer Genome Atlas (TCGA). Indeed, looking for genes whose DNA copy number (the "cause") is significantly associated with their expression level ("the effect") in a cancer study can help pinpoint candidates implied in the disease and improve on our understanding of its molecular bases. DNA methylation (potential confounder) is an important player to account for in this setting, as it can down-regulate gene expression and may also influence DNA copy number.

email: antoine.chambaz@parisdescartes.fr

TARGETED MAXIMUM LIKELIHOOD ESTIMATION: ASSESSING CAUSAL EFFECTS USING HIGH-DIMENSIONAL LONGITUDINAL DATA STRUCTURES

Marco Carone, University of California, Berkeley*
Mark J. van der Laan, University of California, Berkeley

In this talk we present targeted maximum likelihood estimators of a causal effect defined within realistic semiparametric models for the data-generating experiment, eliminating the need to specify parametric regression models. Fundamental components of this methodology include i) the careful definition of the target parameter of the data-generating distribution in a realistic semiparametric model, ii) the aggressive use of cross-validation to select optimal combinations of many candidate estimators, and iii) subsequent targeted maximum likelihood estimation to tailor the fit of the data-generating distribution to the target parameter of interest. Through simulation studies, we demonstrate the performance of this general methodology when interest lies, for example, in assessing effects of single nucleotide polymorphisms in genome-wide case-control studies, or in determining treatment effects in the setting of studies yielding interval-censored time-to-event outcomes and time-dependent covariates allowing for informative dropout.

email: marcocarone@gmail.com

54. BAYES AND OTHER APPROACHES TO VARIABLE AND MODEL SELECTION

DETERMINING ASSOCIATIONS AMONG ENVIRONMENTAL CHEMICALS, NUTRITION AND HEALTH OUTCOMES

Caroline Carr, Virginia Commonwealth University*
Chris Gennings, Virginia Commonwealth University
Roy Sabo, Virginia Commonwealth University
Pam Factor-Litvak, Columbia University

Despite increasing public concern regarding the potential health effects of pervasive environmental exposures, there are few, if any, recommendations for life style changes to mitigate such effects. We are interested in empirically evaluating the complex relationship between environmental exposures and nutrition. Due to the inherent correlations among chemicals and among nutrients, traditional methods are problematic. We propose the formulation of an empirically based weighted index to represent the relative importance of certain chemicals and certain nutrients on a particular health outcome. This method was evaluated through simulations and proves to be less sensitive to complex correlations than standard methods. The importance of the estimated weights was assessed by bootstrapping from the original data. The methods will be presented along with a demonstration using NHANES data. This research was partially supported by NIEHS T32ES007334 and NIH UL1RR031990.

email: ckcarr2487@gmail.com

BAYES VARIABLE SELECTION IN SEMIPARAMETRIC LINEAR MODELS

Suprateek Kundu, University of North Carolina at Chapel Hill*
David B. Dunson, Duke University

There is a rich literature proposing methods and establishing asymptotic properties of Bayesian variable selection methods for parametric models, with a particular focus on the normal linear regression model and an increasing emphasis on settings in which the number of candidate predictors (p) diverges with sample size (n). However often in applications, the residual is found to deviate from normality, thus rendering the above methods ineffective. Our focus is on generalizing methods and asymptotic theory established for mixtures of g-priors to semiparametric linear regression models having unknown residual densities. Using a Dirichlet process location mixture for the residual density, we propose a semiparametric g-prior which incorporates an unknown matrix of cluster allocation indicators. For this class of priors, posterior computation can proceed via a straightforward stochastic search variable selection algorithm. In addition, Bayes factor and variable selection consistency is shown to result under various cases including proper and improper priors on g and $p>n$, with the models under comparison restricted to have model dimensions diverging at a rate less than n.

email: skundu@email.unc.edu

SURE SCREENING FOR ESTIMATING EQUATIONS IN ULTRA-HIGH DIMENSIONS

Sihai D. Zhao, Harvard University*

As the number of possible predictors generated by high-throughput experiments continues to increase, methods are needed to quickly screen out unimportant covariates before fitting regression models. Various screening methods have been proposed and theoretically justified, but so far this has only been done for specific models on a case-by-case basis. In this paper we propose EEScreen, a screening procedure for any model that can be fit using estimating equations, and provide finite-sample performance guarantees. Furthermore, EEScreen requires only a single evaluation of the estimating equation and is more computationally efficient than previous screening methods. We also present an iterative version of EEScreen (iEEScreen), and we show that iEEScreen is closely related to a recently proposed boosting method for estimating equations. We demonstrate our methods on data from a multiple myeloma study, and show via simulations for two different estimating equations that EEScreen and iEEScreen are useful and flexible screening procedures.

email: szhao@hsph.harvard.edu

ESTIMATING LINK FUNCTION PARAMETERS IN ROBUST BAYESIAN BINARY REGRESSION

Vivekananda Roy, Iowa State University*

The logistic and probit regression models are most commonly used to analyze binary response data, but it is well known that their maximum likelihood estimators are not robust to outliers. Liu(2004) proposed a robust regression model, called the robit model, which replaces the normal (logistic) distribution in the probit (logit) regression model with the Student's $t -$distribution. Unlike probit and logistic model, the robit model has an extra degrees of freedom parameter. In this paper, we propose an empirical Bayes approach for estimating the degrees of freedom parameter along with the regression coefficients. We show that a combination of importance sampling based on a fast mixing Markov chain and an application of control variates can be used to efficiently estimate a large class of Bayes factors for selecting the degrees of freedom parameter of the robit model.

email: vroy@iastate.edu

CALIBRATED BAYES FACTORS FOR MODEL COMPARISON

Xinyi Xu, The Ohio State University*
Pingbo Lu, The Ohio State University
Steven MacEachern, The Ohio State University
Ruoxi Xu, The Ohio State University

Bayes factor is a widely used tool for Bayesian hypothesis testing and model comparison. However, it can be greatly affected by the prior elicitation for the model parameters. When the prior information is weak, people often use proper priors with large variances. In this work, we show that when the models under comparisons differ in dimensions, Bayes factors under convenient diffuse priors can be very misleading. Therefore, we propose an innovative method called calibrated Bayes factor, which uses data to calibrate the prior distributions before computing Bayes factors. We show that this method provides reliable and robust model preferences under various true models. It is applicable to a large variety of model comparison problems because it makes no assumption on model forms (parametric or nonparametric) and can be used for both proper and improper priors.

email: xinyi@stat.osu.edu

A SYSTEMATIC SELECTION METHOD FOR THE DEVELOPMENT OF CANCER STAGING SYSTEMS

Yunzhi Lin, University of Wisconsin-Madison*
Richard J. Chappell, University of Wisconsin-Madison
Mithat Gönen, Memorial Sloan-Kettering Cancer Center

The tumor-node-metastasis (TNM) staging system has been the lynchpin of cancer diagnosis, treatment, and prognosis for many years. For meaningful clinical use, an orderly, progressive condensation of the T and N categories into an overall staging system needs to be defined, usually with respect to a time-to-event outcome. This can be considered as a cutpoint selection problem for a censored response partitioned with respect to two ordered categorical covariates and their interaction. The aim is to select the best grouping of the TN categories. A novel bootstrap grouping/model selection method is proposed for this task by maximizing bootstrap estimates of the chosen statistical criteria. The criteria are based on prognostic ability including a landmark measure of area under the ROC curve, a modification of Harrell's c-index, and a concordance probability based on the proportional hazards model. A simulation study was carried out to examine the finite sample performance of the selection procedures based on bootstrapping pairs. We demonstrated that this method is able to give the right grouping with large sample size and that it is not affected by random censorship. We illustrated the utility of our method by applying it to the staging of colorectal cancer.

email: yunzhi@stat.wisc.edu

55. CLUSTERED/REPEATED MEASURES SURVIVAL ANALYSIS

TESTING FOR MONOTONE TIME TREND IN RECURRENT EVENT PROCESSES

Candemir Cigsar, Women's College Research Institute, Princess Margaret Hospital*

A much-studied aspect of processes where individuals or systems experience recurrent events is the existence or non-existence of time-trends. However, a general concept of trend is elusive, and the dependence of the behavior of tests on the assumed definitions of "no trend" and "trend", and on the observation periods for the processes, has been largely ignored. We discuss these issues and study them through analytical results and simulation studies. We also present robust tests for trend across multiple processes, extend them to deal with interval-censored event times, and compare them and other well known trend tests with respect to the issues mentioned. Robust trend tests can also be extended to adjust for time-varying factors such as seasonal effects. Methods are illustrated on a study involving recurrent asthma attacks in children.

email: Candemir.Cigsar@wchospital.ca

CONTRASTING GROUP-SPECIFIC CUMULATIVE MEAN ASSOCIATED WITH MARKED RECURRENT EVENTS IN THE PRESENCE OF A TERMINATING EVENT

Yu Ma, University of Michigan*
Douglas E. Schaubel, University of Michigan

In many biomedical studies where the event of interest is recurrent (e.g., hospital admission), marks are observed upon the occurrence of each event (e.g., medical costs, length of stay). Few papers have been developed under a framework where subjects experience both marked recurrent events and a terminating event (e.g., death), a frequently occurring data structure. We propose novel methods which contrast group-specific cumulative means, influenced by the history of recurrent event and survival experience. Our proposed methods utilize a form of hierarchical modeling: a proportional hazards model for the terminating event; a proportional rates model for the conditional recurrent event rate given survival; and a generalized estimating equations approach for the marks, given an event has occurred. Group-specific cumulative means are estimated (as processes over time) by averaging fitted values from the afore-listed models, the averaging being with respect to the marginal covariate distribution. Large sample properties are derived, while simulation studies are conducted to assess finite sample properties. We apply the proposed methods to data obtained from the CANADA-USA Peritoneal Dialysis Study (CANUSA), which motivated our research.

email: rickma@umich.edu

ALTERNATING EVENT PROCESSES DURING LIFETIMES: POPULATION DYNAMICS AND STATISTICAL INFERENCE

Russell T. Shinohara, Johns Hopkins University*
Mei-Cheng Wang, Johns Hopkins University

In the literature studying recurrent event data, a large amount of work has been focused on univariate recurrent event processes in which the occurrence of each event is treated as a single point in time. There are many applications, however, in which patients experience nontrivial durations associated with each event. This results in a process where the disease status of a patient alternates between exacerbations and remissions. In this paper, we consider the dynamics of a chronic disease and its associated exacerbation-remission process over two time scales: calendar time and time-since-onset. In particular, over calendar time, we explore population dynamics and the relationship between incidence, prevalence and duration for such alternating event processes. We provide nonparametric estimation techniques for characteristic quantities of the process. In many settings, exacerbation processes are observed from an onset time until death; this induces informative censoring. Using a nonparametric latent variable approach to account for the relationship between the survival and alternating event processes, we develop techniques for estimating semiparametric models of prevalence. By combining population dynamics and within-process structure, the proposed approaches provide an alternative and general way to study alternating event processes.

email: taki.shinohara@gmail.com

SEMIPARAMETRIC PROBIT MODEL FOR CLUSTERED INTERVAL-CENSORED DATA WITH UNKNOWN DISTRIBUTION OF RANDOM EFFECTS

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

Clustered interval-censored data is commonly arised in many follow-up medical studies. Incorporating the cluster effect will improve the estimation efficiency. In this paper, we propose frailty Probit model for analyzing such data. We allow the distribution of the frailties to be unknown by assigning it a Dirichlet Process mixture prior. An efficient and easy to implement Gibbs sampler is proposed for the estimation. Our method is evaluated by a simulation study and illustrated by an application to a infectious disease data.

email: wuh@email.sc.edu

A FLEXIBLE COPULA MODEL FOR BIVARIATE SURVIVAL DATA

Zhen Chen, University of Rochester Medical Center*
David Oakes, University of Rochester Medical Center
Ollivier Hyrien, University of Rochester Medical Center
Changyong Feng, University of Rochester Medical Center

Copulas are bivariate distributions with uniform marginals. They provide a general method for linking univariate marginal distributions to form multivariate distribution. Following Clayton (1978), several families of single-parameter copula models have been proposed for analyzing survival data. This article explores the use of a flexible two-parameter copulas family for bivariate survival data. The basic properties of this family are described. Under parametric assumptions on the univariate marginals, besides the commonly used one-stage M.L.E., a two-stage parameter estimation method is developed. In addition, we generalize the two-stage estimation approach for the semi-parametric model when the distributions of the univariate marginals are unspecified. The asymptotic properties of all proposed estimators are derived and compared by simulations under different censorship scenarios. We find that the one-stage and two-stage parametric estimators perform similarly, with somewhat lower variances than the two-stage semi-parametric estimator. The practical application of the model is illustrated through a data example.

email: zhen_chen@urmc.rochester.edu

56. GENOMICS

THE PRACTICAL EFFECT OF BATCH ON PREDICTION

Hilary S. Parker, Johns Hopkins School of Public Health*
Jeffrey T. Leek, Johns Hopkins School of Public Health

Measurements from microarray and other high-throughput technologies are susceptible to a number of biological and non-biological artifacts like batch effects. It is known that batch effects can alter or obscure the set of significant results and biological conclusions in high-throughput experiments. Here we examine the impact of batch effects on predictors built from genomic technologies. To investigate batch effects, we collected publicly available gene expression measurements with known outcomes and batches. Using these data we show: (1) the impact of batch effects on prediction depends on the correlation between outcome and batch in the training data, (2) removing expression measurements most affected by batch before building predictors may improve the accuracy of those predictors. These results suggest that (1) training sets should be designed to minimize correlation between batches and outcome, (2) methods for identifying batch-affected probes should be developed to improve prediction results for studies with high correlation between batches and outcome.

email: hiparker@jhsph.edu

IDENTIFYING AND CORRECTING SAMPLE MIX-UPS IN HIGH-DIMENSIONAL DATA

Karl W. Broman, University of Wisconsin-Madison*
Mark P. Keller, University of Wisconsin-Madison
Aimee T. Broman, University of Wisconsin-Madison
Danielle M. Greenawalt, Merck & Co., Inc.
Christina Kendziorski, University of Wisconsin-Madison
Eric E. Schadt, Pacific Biosciences
Saunak Sen, University of California, San Francisco
Brian S. Yandell, University of Wisconsin-Madison
Alan D. Attie, University of Wisconsin-Madison

In a mouse intercross with more than 500 animals and genome-wide gene expression data on six tissues, we identified a high proportion of sample mix-ups in the genotype data, on the order of 15%. Local eQTL (genetic loci influencing gene expression) with extremely large effect may be used to form a classifier for predicting an individual's eQTL genotype from its gene expression value. By considering multiple eQTL and their related transcripts, we identified numerous individuals whose predicted eQTL genotypes (based on their expression data) did not match their observed genotypes, and then went on to identify other individuals whose genotypes did match the predicted eQTL genotypes. The concordance of predictions across six tissues indicated that the problem was due to mix-ups in the genotypes. Consideration of the plate positions of the samples indicated a number of off-by-one and off-by-two errors, likely the result of pipetting errors. Such sample mix-ups can be a problem in any genetic study. As we show, eQTL data allow us to identify, and even correct, such problems.

email: kbroman@biostat.wisc.edu

DETECTING DIFFERENTIAL BINDING OF TRANSCRIPTION FACTORS WITH ChIP-Seq

Kun Liang, University of Wisconsin-Madison*
Sunduz Keles, University of Wisconsin-Madison

Increasing number of ChIP-seq experiments are investigating transcription factor binding under multiple experimental conditions, for example, various treatment conditions, several distinct time points, and different treatment dosage levels. Hence, identifying differential binding sites across multiple conditions is of practical importance in biological and medical research. To this end, we have developed a statistical method to detect differentially bound sharp binding sites across multiple conditions, with or without matching control samples. Application of our method can lead to meaningful downstream analyses and elucidation of the functional roles of transcription factors across different conditions.

email: liangkun1@gmail.com

APPLYING WHOLE GENOMIC PREDICTION ACROSS POPULATIONS FOR PREDICTIVE AND PROGNOSTIC PURPOSES

Robert Makowsky, U.S. Food and Drug Administration*
Kirk Yancy B. Williams, U.S. Food and Drug Administration
Gustavo de los Campos, University of Alabama at Birmingham

Whole Genome Prediction (WGP) offers an increased predictive accuracy compared to conventional single marker approaches. However, the methods have only been employed in humans sparingly and it is not known how population structure will impede accurate predictions and, if so, how best to cope with such a situation. Using Forward-Time Simulations as employed by FREGENE, we produced genomic data from three human populations, where gene flow ceased 15-50 thousand years ago. This data is designed to mimic the population history of humans. Genomic values were calculated using 1,000 loci and individual marker effects drawn from a double-exponential distribution with the narrow-sense heritability ranging from 0.25-0.75. Genomic data mimicking a Single Nucleotide Polymorphism (SNP) chip were used to predict phenotypes of independent individuals while varying each population's representation in the training and testing datasets. We find that predictive ability is reduced and offer insights into how best to cope with datasets representing individuals of different ancestries.

email: lokido@uab.edu

SEGMENTING THE HUMAN GENOME BASED ON MUTATION RATES

Prabhani Kuruppumullage Don, Pennsylvania State University*
Guruprasad Ananda, Pennsylvania State University
Francesca Chiaromonte, Pennsylvania State University
Kateryna D. Makova, Pennsylvania State University

Various studies have used Hidden Markov Models (HMMs) to segment the human genome based on sequence patterns. However, none to date has attempted a segmentation based on the large scale behavior of mutation rates , something that can provide crucial insights into genome dynamics and evolution. Here, we partitioned each of the 22 human autosomal chromosomes into 1-Mb windows and obtained the rates of nucleotide substitutions, small insertions and deletions (indels), and mononucleotide microsatellite repeat number alterations in these windows separately from two sub-genomes representative of neutral evolution; ancestral repeats (AR) and non-coding, non-repetitive sequences (NCNR). We applied Multivariate Gaussian HMMs to identify the underlying states of neutral variation defined jointly by these rates, and to segment the human genome accordingly. Both AR and NCNR rates delineated biologically meaningful states representing hot and cold regions (increased or decreased indels and substitutions). Cold regions comprise long segments positioned in the middle ranges of chromosomes, while hot regions comprise medium-sized segments largely occurring in sub-telomeric regions. Notably, both AR and NCNR rates also delineated a distinct state with markedly increased microsatellite mutability rates, which comprises very short segments interspersed along the chromosomes.

email: pxk919@psu.edu

IDENTIFYING PROTEIN BINDING SITES FROM GENOMIC ChIP-Seq DATA USING REVERSIBLE JUMP MCMC

Rasika V. Jayatillake, Old Dominion University*
Nak-Kyeong Kim, Old Dominion University

Chromatin Immunoprecipitation followed by massively parallel sequencing (ChIP-seq) is a new technology that can reveal protein binding sites in genome with superior accuracy. Although there are many models that can estimate binding sites with some level of success, most of them use simple moving window based method or normal density kernels. Moreover, when there are two or more binding sites in a short distance, many of the existing models cannot predict those multiple sites. In this study, a Bayesian model is considered to predict multiple binding sites in short distances. Since the number of binding events in a region is unknown, a reversible jump MCMC is proposed. The results on simulated data and ChIP-seq data for transcription factor STAT1 on human genome will be presented.

email: rjayatil@odu.edu

CHANGE-POINT ANALYSIS OF PAIRED ALLELE-SPECIFIC COPY NUMBER VARIATION DATA

Yinglei Lai, The George Washington University*

The recent genome-wide allele-specific copy number variation data enable us to explore two types of genomic information including chromosomal genotype variations as well as DNA copy number variations. For a cancer study, it is common to collect data for paired normal and tumor samples. Then, two types of paired data can be obtained to study a disease subject. However, there is a lack of methods for a simultaneous analysis of these four sequences of data. In this study, we propose a statistical framework based on the change-point analysis approach. The validity and usefulness of our proposed statistical framework are demonstrated through the simulation studies and applications based on an experimental data set.

email: ylai@gwu.edu

57. HEALTH SERVICES/HEALTH POLICY

IDENTIFYING INDIVIDUAL CHANGES IN PERFORMANCE WITH COMPOSITE QUALITY INDICATORS WHILE ACCOUNTING FOR REGRESSION-TO-THE-MEAN

Byron Gajewski, University of Kansas*
Nancy Dunton, University of Kansas

Almost a decade ago Morton & Torgerson (p. 1084) indicated that perceived medical benefits could be due to "regression-to-the-mean." Despite this caution, the "regression-to-the-mean" effects on the identification of changes in institutional performance do not seem to have been considered previously in any depth (Jones & Spiegelhalter, p. 1646). As a response, Jones & Spiegelhalter provide a methodology to adjust for regression-to-the-mean when modeling recent changes in institutional performance for one variable quality indicators (QIs). Therefore, in our view, Jones & Spiegelhalter provide a breakthrough methodology for performance measures. At the same time, in the interests of parsimony, it is useful to aggregate individual QIs into a composite score. Our question is: Can we develop and demonstrate a methodology that extends the "regression-to-the-mean" literature to allow for composite quality indicators? Using a latent variable modeling approach, we extend the methodology to the composite indicator case. We demonstrate the approach on four indicators collected by the National Database of Nursing Quality Indicators® (NDNQI®). A simulation study further demonstrates its "proof of concept."

email: bgajewski@kumc.edu

ALCOHOL OUTLETS AND VIOLENCE IN THE CITY OF PHILADELPHIA: THE ROLE OF LAND USE

Tony H. Grubesic, Drexel University
Loni Philip Tabb, Drexel University
Dominique Williams, Drexel University*
William Pridemore, Indiana University-Bloomington

The relationship between alcohol outlet density and violence in urban environments is complex. Although the local ecological characteristics of neighborhoods, such as socio-economic status, demographic composition and community organization play a role in the outlet-violence connection, far less is known about the potential moderating effects of land use on violence. The purpose of this paper is to use a suite of cross-sectional regression models to explore the impacts of land use on violence in the city of Philadelphia. Specifically, while previous work addresses land use in aggregate, typically at the block group, tract or ZIP code level, this paper will explore land use in direct proximity to alcohol outlets using a series of network-based catchment areas. Policy implications are discussed.

email: grubesic@drexel.edu

CALIBRATED SENSITIVITY ANALYSIS FOR THE INSTRUMENTAL VARIABLES METHOD FOR OBSERVATIONAL STUDIES

Jesse Yenchih Hsu, University of Pennsylvania*
Scott A. Lorch, University of Pennsylvania
Dylan S. Small, University of Pennsylvania

The instrumental variables (IV) method is an approach to estimating a causal relationship between a treatment and an outcome based on an observational study in the presence of unmeasured confounding variables. A valid IV is a variable that affects the treatment, has no direct effect on the outcome other than through its effect on the treatment, and is independent of the unmeasured confounders given measured confounders. There is often concern that an IV is not perfectly valid in the sense that it is correlated with unmeasured confounders. A sensitivity analysis is used to examine the impact of the violation. In many available sensitivity analysis methods, the sensitivity parameter that describes the invalidity of the proposed IV is on an absolute scale. It may be difficult for a subject matter expert to specify a plausible range of values for the sensitivity parameter on this absolute scale. We develop an approach that calibrates the value of the sensitivity parameter to the observed covariates that are thought to be related to the unmeasured confounders that are making the proposed IV invalid and to be more interpretable to subject matter experts. We will illustrate our method using a neonatology study.

email: hsu9@wharton.upenn.edu

A RESEARCH AGENDA: DOES GEOCODING POSITIONAL ERROR MATTER IN HEALTH GIS STUDIES?

Geoffrey M. Jacquez, BioMedware*

Geocoding positional errors can impact disease rates; disease clustering; odds ratios; and estimates of individual exposures, resulting in flawed findings and poor public health decisions. Yet, geocoding positional error is often ignored, primarily because of a lack of theory, methods and tools for propagating positional errors. Important knowledge gaps include a detailed understanding of empirical geocoding error distributions and their spatial autocorrelation structure; impacts of positional error on the statistical power of spatial analysis methods; models for predicting positional error at specific locations; and the propagation of positional errors through health analyses. A research agenda is proposed to address five key needs. A lack of: 1) Standardized, geocoding resources for use in health research; 2) Validation datasets that will allow the evaluation of alternative geocoding procedures; 3) Spatially explicit geocoding positional error models; 4) Resources for assessing the sensitivity of spatial analysis results to positional error; 5) Demonstration studies of the sensitivity of health policy decisions to geocoding positional accuracy.

e-mail: jacquez@biomedware.com

METHODOLOGY FOR SCORING THE EQ-5D

Eleanor M. Pullenayegum, McMaster University*
Feng Xie, McMaster University

Health utilities are the quality weights used to calculate quality adjusted life years, a common outcome in health research. In most studies, health utilities are measured using standardised questionnaires, and one of the most popular choices is the EQ-5D. A new version of the EQ-5D, offering more precise measurement, is currently being developed. The questionnaire consists of five questions, to each of which respondents must choose one of five responses, yielding 3125 possible health states. These health states must be converted into utilities through a scoring algorithm. Given the large number of health states in the new EQ-5D, methodology for creating the scoring algorithm is lacking. The data that will be used to calibrate the scoring algorithm consist primarily of discrete choice experiments, with a smaller number of time-trade-off tasks. This suggests using Generalised Linear Mixed Models to construct latent utilities, and regression techniques to map latent utilities onto utilities. Through a series of simulation studies, we explore the performance of these methods under varying inter-rater agreement and sample sizes. Given that the true within-individual correlation structures are unknown, and that it is impractical to evaluate all possible interaction terms, we also examine the robustness to model mis-specification.

e-mail: pullena@mcmaster.ca

ESTIMATING 95% CONFIDENCE INTERVAL FOR PERCENTILE RANK – USING BOOTSTRAP –APPLICATION: RSMR & RSRR

Yahya A. Daoud, Baylor Health Care System*
Yumi Y. Sembongi, Baylor Health Care System
Monica Anand, Baylor Health Care System
Dunlei Cheng, Baylor Health Care System
Edward B. De Vol, Baylor Health Care System

As consulting biostatisticians within a heath care system, we are often faced with unique and challenging problems. Baylor Health Care System adopted the percentile ranks (PR) as a method to report our performance compared with national data. Variation in these rankings over time and between hospitals have prompted system leaders to ask if the changes in the percentile ranks are statistically significant. We decided that utilizing the 95% confidence interval (CI) would help answer this question and provide a better understanding of the performance for each measure. Because there was not a well established method for determining the required 95% confidence interval, we developed a generalized bootstrap simulation program using SAS to estimate the required confidence interval. We utilized CMS Hospital Compare databases to evaluate the overall BHCS performance as well as hospital level performance for 10 BHCS hospitals for the periods 2005-2008 and 2006-2009 by estimating the 95% CI for the metrics 30 Days Post-admission Risk Standardized Mortality Ratio (RSMR) and 30 Days Post-admission Risk Standardized Readmission Ratio (RSRR) for three conditions: Acute Myocardial Infarction (AMI), Heart Failure (HF or CHF) and Pneumonia (PNE). The distributions of the bootstrap PR data were normally distributed for almost all PR except at 100% PR.

e-mail: YahyaD@BaylorHealth.edu

OPTIMIZATION AND SIMULATION OF AN EVOLVING KIDNEY PAIRED DONATION (KPD) PROGRAM

Yijiang J. Li, University of Michigan*
Peter X. K. Song, University of Michigan
Yan Zhou, University of Michigan
Alan B. Leichtman, University of Michigan
Michael A. Rees, University of Toledo Medical Center
John D. Kalbfleisch, University of Michigan

Kidney paired donation (KPD) programs provide a unique and important platform for living incompatible donor-candidate pairs to exchange organs in order to achieve mutual benefit. We propose a novel approach to organizing kidney exchanges in an evolving KPD program with advantages, including (1) allowance for a more general medical-outcome-based evaluation of potential kidney transplants; (2) consideration of stochastic features in managing a KPD program; and (3) exploitation of possible alternative exchanges when the originally planned allocation cannot be fully executed. Another primary contribution of this work is rooted in the development of a comprehensive microsimulation system for simulating and studying various aspects of an evolving KPD program. Three allocation strategies are proposed and obtained based on an integer programming (IP) formulation, and microsimulation models can allow tracking of the evolving KPD program over a series of match runs to evaluate different allocation strategies. Simulation studies are provided to illustrate our proposed methods.

e-mail: yijiang@umich.edu

58. TOWARDS OMICS-BASED PREDICTORS FOR PATIENT MANAGEMENT

VALIDATING CLINICAL PERFORMANCE OF PREDICTORS

Michael L. LeBlanc, Fred Hutchinson Cancer Research Center*

We discuss strategies to establish adequate performance of a predictor based on trial data collected retrospectively prior to use in a clinical trial. Internal validation (or cross-validation) can be a useful tool in model selection and in obtaining less biased prediction error estimates. However, a fully separate validation sample is the preferred method to obtain an unbiased assessment of performance. While statistically straight-forward, a convincing implementation of this validation strategy can be challenging in practice, especially when collaborations involve multiple research groups. Issues discussed include: the importance of a fully specified prospective analysis protocol and algorithm, blinded assessment, assay or sample consistency between training and validation datasets and case selection.

email: mleblanc@fhcrc.org

A REGULATORY PERSPECTIVE ON OMICS-BASED PREDICTORS

Gene A. Pennello, U.S. Food and Drug Administration*

An omics-based predictor is an example of a diagnostic medical device or test. At the Food and Drug Administration (FDA), diagnostic tests submitted for pre-market approval or clearance are reviewed by the Center for Devices and Radiological Health (CDRH). CDRH employs a risk-based approach to determine the regulatory pathway of medical devices. For example, a prognostic test may be determined to be of moderate risk, while a companion diagnostic test, i.e., a test that is essential for safe and effective use of a therapeutic product, is often determined to be of high risk. In this talk, I will discuss FDAregulation of diagnostic tests, provide examples of omics-based predictors that have undergone regulatory review, and discuss challenges with the clinical and analytical validation of such tests.

email: gene.pennello@fda.hhs.gov

STATISTICAL ISSUES IN THE DESIGN OF CLINICAL TRIALS TO ESTABLISH THE UTILITY OF BIOMARKER-BASED TESTS FOR GUIDING THERAPY DECISIONS

Lisa M. McShane, National Cancer Institute, National Institutes of Health*

A better understanding of disease through identification of biological characteristics of the disease process and host that predict disease course and responsiveness to therapy will be essential for the discovery of new therapies and for optimization of clinical care for individual patients. There are many ongoing efforts to develop biomarker-based tests that could provide these clinically informative biological characterizations. Tests that will influence therapy decisions require definitive evaluation of their clinical utility in appropriately designed clinical trials, just as novel treatments require rigorous evaluation. It should be recognized that the statistical requirements for establishing utility of a biomarker-based test are different than those for establishing treatment benefit. Key considerations are the need for appropriate control groups to avoid the potential for confounding of prognostic and predictive effects, and the choice of appropriate metrics and statistical tests to establish biomarker-based test performance and clinical value. We discuss a variety of trial designs that have been proposed for evaluation of biomarker-based tests and provide an in-depth comparison of their advantages and disadvantages.

email: lm5h@nih.gov

59. FUNCTIONAL DATA ANALYSIS

METHODOLOGY AND THEORY FOR PARTIAL LEAST SQUARES APPLIED TO FUNCTIONAL DATA

Peter Hall, The University of Melbourne and University of California, Davis*
Aurore Delaigle, The University of Melbourne

The partial least squares method was originally developed to estimate slope parameters in multivariate parametric models. More recently it has gained popularity for the analysis of functional data literature. There, the partial least squares estimator of slope is used either to construct linear predictive models, or as a tool to project the data onto a one-dimensional quantity that is employed for further statistical analysis. Although the partial least squares approach is often viewed as an attractive alternative to projections onto the principal component basis, its properties are less well known than those of the latter, mainly because of its iterative nature. In this talk we develop an explicit formulation of partial least squares for functional data, which leads to insightful results and motivates new theory, demonstrating consistency and establishing convergence rates.

email: halpstat@ms.unimelb.edu.au

TIME-DYNAMIC FUNCTIONAL ADDITIVE MODEL

Jane-Ling Wang, University of California at Davis*
Xiaoke Zhang, University of California at Davis
Byeong Park, Seoul National University

Additive model is an effective dimension reduction model that provides flexibility to model the relation between a response variable and key covariates. The literature is largely developed for scalar response and vector covariates. In this paper, more complex data is of interest, where both the response and covariates may be functions. A functional additive model is proposed together with a smooth backfitting algorithm to estimate the unknown regression functions, whose components are time-dependent additive functions of the covariates. Due to the sampling plan, such functional data may not be completely observed as measurements may only be collected intermittently at discrete time points. We develop a uniform platform and efficient approach that can cover both dense and sparse functional data and the needed theory for statistical inference. The oracle properties of the component functions are also established.

email: jlwang.ucdavis@gmail.com

CONTINUOUSLY ADDITIVE MODELS FOR FUNCTIONAL REGRESSION

Hans-Georg Mueller, University of California at Davis
Yichao Wu, North Carolina State University*
Fang Yao, University of Toronto

We propose Continuously Additive Models (CAM), an extension of additive regression models to the case of infinite-dimensional predictors, corresponding to smooth random trajectories, coupled with scalar responses. As the number of predictor times and thus the dimension of predictor vectors grows larger, properly scaled additive models for these high-dimensional vectors are shown to converge to a limit model, in which the additivity is conveyed through an integral. This defines a new type of functional regression model. In these Continuously Additive Models, the path integrals over paths defined by the graphs of the functional predictors with respect to a smooth additive surface relate the predictor functions to the responses. This is an extension of the situation for traditional additive models, where the values of the additive functions, evaluated at the predictor levels, determine the predicted response. We study prediction in this model, using tensor product basis expansions to estimate the smooth additive surface that characterizes the model. In a theoretical investigation, we show that the predictions obtained from fitting continuously additive estimators are asymptotically consistent. We also consider extensions to generalized responses.

email: wu@stat.ncsu.edu

MOVELETS: A DICTIONARY OF MOVEMENT

Bai Jiawei, Johns Hopkins University
Jeffrey Goldsmith, Johns Hopkins University
Ciprian M. Crainiceanu, Johns Hopkins University*

Recent technological advances provide researchers a way of gathering real-time information on an individual's movement through the use of wearable devices that record acceleration. In this paper, we propose a method for identifying activity types, like walking, standing, and resting, from acceleration data. Our approach decomposes movements into short components called "movelets", and builds a reference for each activity type. Unknown activities are predicted by matching new movelets to the reference. We apply our method to data collected from a single, three-axis accelerometer and focus on activities of interest in studying physical function in elderly populations. An important technical advantage of our methods is that they allow identification of short activities, such as taking two or three steps and then stopping, as well as low frequency rare activities, such as sitting on a chair. Based on our results we provide simple and actionable recommendations for the design and implementation of large epidemiological studies that could collect accelerometry data for the purpose of predicting the time series of activities and connecting it to health outcomes.

email: ccrainic@jhsph.edu

60. THE ANALYSIS OF SOCIAL NETWORK DATA IN PUBLIC HEALTH

NETWORK BASED METHODS FOR ACCESSING HARD-TO-REACH GROUPS

Tyler H. McCormick, University of Washington*
Tian Zheng, Columbia University

The sampling frame in most social science surveys excludes members of certain groups, known as hard-to-reach groups. These groups may be be difficult to access (the homeless, for example), camouflaged by stigma (individuals with HIV/AIDS), or both (commercial sex workers). We develop a Bayesian Hierarchical model which leverages social structure in respondents' social networks to access these individuals. The data used in this work measure social network structure indirectly through questions on standard surveys and do not require a special sampling scheme. This model estimates relative homogeneity between groups in the population and variation in the propensity for interaction between respondents and group members. The model also estimates features of groups which are difficult to reach using standard surveys.

email: tylermc@u.washington.edu

USING RETROSPECTIVE SAMPLING TO STUDY FACTORS AFFECTING RELATIONSHIPS IN LARGE LONGITUDINAL SOCIAL NETWORKS

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

Statistical modeling of longitudinal sociocentric data (values describing the relationship between all pairs of individuals) is complicated by the complex dependencies in the data and the quadratic growth in the number of observations with the number of individuals. In this talk we exploit: 1) the longitudinality of the data and 2) the natural phenomena that large social networks are dominated by null relationships to propose a model and estimation approach that enables feasible computations on very large networks. The key characteristics of the model is that it accounts for dependencies between pairs of actors (dyads) using lagged observed predictors and contemporaneous latent variables, thus retaining a form of conditional independence between dyads. The novel feature of the estimation approach is that we fit the model to a randomly selected subset of dyads and use retrospective sampling weights to recover consistent estimates of the model parameters. Theoretical properties of sampling schemes such as sampling only always-null dyads and the impact of the sampling probabilities on the efficiency of estimates are discussed using real data from a large social network.

email: omalley@hcp.med.harvard.edu

POINT PROCESS MODELING FOR DIRECTED INTERACTION NETWORKS

Patrick O. Perry, New York University*
Patrick J. Wolfe, Harvard University

Network data often take the form of repeated interactions between senders and receivers tabulated over time. A primary question to ask of such data is which traits and behaviors are predictive of interaction. To answer this question, a model is introduced for treating directed interactions as a multivariate point process: a Cox multiplicative intensity model using covariates that depend on the history of the process. Consistency and asymptotic normality are proved for the resulting partial-likelihood-based estimators under suitable regularity conditions, and an efficient fitting procedure is described. Multicast interactions--those involving a single sender but multiple receivers--are treated explicitly. The resulting inferential framework is then employed to model message sending behavior in a corporate e-mail network. The analysis gives a precise quantification of which static shared traits and dynamic network effects are predictive of message recipient selection.

email: pperry@stern.nyu.edu

61. NOVEL METHODOLOGICAL ISSUES IN ANALYZING AND DESIGNING LONGITUDINAL BIOMARKER STUDIES

OUTCOME DEPENDENT SAMPLING FOR LONGITUDINAL BINARY RESPONSE DATA BASED ON A TIME-VARYING AUXILIARY VARIABLE

Jonathan S. Schildcrout, Vanderbilt University*
Sunni L. Mumford, 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
Patrick J. Heagerty, University of Washington
Paul J. Rathouz, University of Wisconsin

Outcome dependent sampling (ODS) study designs are commonly implemented with rare diseases or when prospective studies are infeasible. In longitudinal data settings, when a repeatedly measured binary response is rare, an ODS design can be highly efficient for maximizing statistical information subject to resource limitations that prohibit covariate ascertainment of all observations. We an ODS design where individual observations are sampled with probabilities determined by an inexpensive, time-varying auxiliary variable that is related but is not equal to the response. With the goal of validly estimating marginal model parameters based on the resulting biased sample, we propose a semi-parametric, Sequential Offsetted Logistic Regressions (SOLR) approach. Motivated by an analysis of the BioCycle Study (Gaskins et al, 2009) that aims to describe the relationship between reproductive health (determined by luteinizing hormone levels) and fiber consumption, we examine properties of SOLR estimators and compare them to other common approaches.

email: jonathan.schildcrout@vanderbilt.edu

A PRINCIPAL INTERACTIONS ANALYSIS FRAMEWORK FOR REPEATED MEASURES DATA ON QUANTITATIVE TRAITS: APPLICATION TO LONGITUDINAL STUDIES OF GENE-ENVIRONMENT INTERACTIONS

Bhramar Mukherjee, University of Michigan*
Yi-An Ko, University of Michigan

Many existing cohorts with longitudinal data on environmental exposures, occupational history, lifestyle/behavioral characteristics and health outcomes have collected genetic data in recent years. We consider the problem of modeling gene-environment interactions with repeated measures data of a quantitative trait and time varying exposure. We explore a class of interaction models that are based on a singular value decomposition of the cell means residual matrix after fitting the additive main effect terms. This class of additive main effects and multiplicative interaction (AMMI) models (Gollob, 1968) provide useful summaries for subject-specific and time-varying effects as represented in terms of their contribution to the leading principal

components/eigenroot of the interaction matrix. It also makes the interaction structure more amenable to geometric representation. We call this analysis "Principal Interactions Analysis": (PIA). The proposed methods are illustrated by using data from the Normative Aging Study, a longitudinal cohort study of Boston area veterans since 1963.

email: bhramar@umich.edu

POOLING DESIGNS FOR OUTCOMES UNDER A GAUSSIAN RANDOM EFFECTS MODEL

Yaakov Malinovsky, University of Maryland, Baltimore County*
Paul S. Albert, 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

Due to the rising cost of laboratory assays, it has become increasingly common in epidemiological studies to pool biospecimens. This is particularly true in longitudinal studies, where the cost of performing multiple assays over time can be prohibitive. In this work, we consider the problem of estimating the parameters of a Gaussian random effects model when the repeated outcome is subject to pooling. We consider different pooling designs for the efficient maximum-likelihood estimation of variance components, with particular attention to estimating the intraclass correlation coefficient (ICC). We evaluate the efficiencies of different pooling design strategies using analytic and simulation study results. We examine the robustness of the designs to skewed distributions and consider unbalanced designs. The design methodology is illustrated with a longitudinal study of pre-menopausal women focusing on assessing the reproducibility of F2-isoprostane, a biomarker of oxidative stress, over the menstrual cycle.

email: yaakovm@umbc.edu

A BAYESIAN ORDER RESTRICTED MODEL FOR HORMONAL DYNAMICS DURING MENSTRUAL CYCLES OF HEALTHY WOMEN

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

We propose a Bayesian framework for analyzing multivariate linear mixed effect models with linear constraints on the fixed effect parameters. Two appraoches, one based on Minkowski-Weyl priors on polygonal regions and another based on projections via quadratic programming are explored. The procedure can incorporate both firm and soft restrictions on the parameters and Bayesian model selection for the random effects. The framework is used to analyze data from the BioCycle Study. One of the main objectives of the BioCycle study is to investigate the association between markers of oxidative stress and hormone levels during menstrual cycles of healthy women. Contrary to the popular belief that ovarian hormones, especially estrogen levels, are negatively associated with level of F2-isoprostanes, a known marker for oxidative stress, our analysis finds a positive association between estrogen and isoprostane levels. The positive association corroborates the findings from a previous analysis of the BioCycle data.

email: anindya@umbc.edu

62. ADVANCES IN CANCER RISK PREDICTION MODELS

MODEL VALIDATION AND UPDATING

Ewout W. Steyerberg, Erasmus University Medical Center, Rotterdam, the Netherlands*

Prediction models are becoming increasingly popular in medicine. There is a widely agreed need for validation before predictions from such models should be applied in medical practice. Validation may show a disappointing predictive performance, suggesting a need to update a model. Updating may consider simple adjustments such as recalibration, but also extension with promising new markers for diagnosis or prognosis. In this talk, potential approaches will be discussed and illustrated with several case studies. The general viewpoint is that prediction models should be developed, validated, and updated in an interative process rather than seen as fixed entities or developed from scratch for each new setting.

email: e.steyerberg@erasmusmc.nl

DEPLOYING STATISTICAL PREDICTION MODELS

Michael W. Kattan, Cleveland Clinic*

A new website will be demonstrated, http://makercalc.ccf.org. This site allows a statistician to very quickly post a prediction model for others (e.g., clinicians) to run. The statistician user simply pastes in his equation, answers a few housekeeping questions (e.g., polite variable names) and hits Submit. He then receives a URL to post on his own website and/or directly distribute to users. The user can then click on the link with his/her desktop computer or handheld device to run the prediction model. The website is under continuous development, and suggestions for improvement are very much appreciated."

email: kattanm@ccf.org

ON JOINT RISK PREDICTION

Ruth Pfeiffer, National Cancer Institute, National Institutes of Health*

Breast, endometrial and ovarian cancers share some hormonal and epidemiologic risk factors. While several models predict absolute risk of breast cancer there are few models for ovarian cancer in the general population and none for endometrial cancer. We recently developed three models to predict absolute risk for these cancers. Using data on white, non-Hispanic women ages 50+ years from two large prospective population-based cohorts, we estimated relative- and attributable-risks for the 3 cancers and combined these estimates with baseline age-specific SEER incidence and competing mortality rates. For endometrial cancer, we allowed for the possibility of hysterectomy during the projection interval by treating hysterectomy as a competing risk and likewise, we allowed for the possibility of oophorectomy for the ovarian cancer model. The models were validated using independent data from the Nurses' Health Study cohort. The uses of the models in medical decision-making or in studies of interventions that impact risk of two or more of these cancers are illustrated by computing the number of endometrial or ovarian cancer cases that fall into the highest percent of risk for breast cancer.

email: pfeiffer@mail.nih.gov

DYNAMIC PREDICTION: UPDATING MEDICAL PREDICTION MODELS IN REAL TIME USING ROUTINELY COLLECTED CLINICAL DATA (OR: WHY CAN'T NOMOGRAMS BE MORE LIKE NETFLIX?)

Andrew J. Vickers, Memorial Sloan-Kettering Cancer Center*

Medical prediction models are often considered to be scientific facts, true for all people, forever and always. But medical practice is highly dynamic. For example, the Kattan nomogram predicts recurrence after surgery for prostate cancer using information on stage, grade and PSA on the basis of patients treated by a single surgeon in the 1990's. Since that time there has been a stage shift, changes in the approach to grading, and improvements in surgery that reduce recurrence risk. In this paper, I argue that medical prediction should follow the approach taken by the most high profile prediction system in the US, that used by Netflix. Netflix asks users to rate a number of movies from 1 to 5 stars and then predicts a star rating for that user for all 35,000 movies in its database. The prediction algorithm, based on machine learning, is continually updated as new ratings become available. I will describe a comparable medical prediction system. R statistical code has been directly incorporated in the electronic health record that updates prediction models in real time as new data become available. These models are then evaluated by directly comparing predictions for a given patient with actual outcome followed prospectively.

email: vickersa@mskcc.org

63. ADAPTIVE DESIGN IN VACCINE TRIALS

A 2-STAGE ADAPTIVE DESIGN FOR ASSESSING VACCINE EFFICACY WITH UNCERTAIN INCIDENCE RATE

Ivan SF Chan, Merck Research Laboratories*
Xiaoming Li, Gilead Sciences
Keaven M. Anderson, Merck Research Laboratories

In many vaccine efficacy studies where the endpoint is a rare infection/disease event, an event-driven design (conditional on the total number of events) is commonly used for testing the hypothesis that study vaccine lowers the risk of the event. Uncertainty of the incidence rate has a large impact on the sample size and study duration. To mitigate the risk of running a potentially large, long-duration efficacy trial with an uncertain event rate, we propose a two-stage adaptive design strategy with interim analyses to allow evaluation of study feasibility and sample size adaptation. During Stage I, a small number of subjects will be enrolled and the feasibility of the study will be evaluated based on the incidence rate observed. If the feasibility of the study is established, at the end of Stage I an interim analysis will be performed with a potential sample size adaptation based the conditional rejection probability approach. The operating characteristics of this design are evaluated by simulation.

email: Ivan_Chan@Merck.Com

ADAPTIVE DESIGNS FOR VACCINE CLINICAL TRIALS

Ghideon Ghebregiorgis, U.S. Food and Drug Administration*

Despite the increase of our understanding of host immune responses, the sequencing of pathogen genomes, and other technological advances, important hurdles remain for developing vaccines for a variety of diseases. Developing a vaccine against diseases like the human immunodeficiency virus (HIV), tuberculosis (TB), Malaria etc poses a challenge in vaccine clinical trial studies. Adaptive clinical trial designs allow a trial to be modified in response to data acquired during the study. Such trials would rapidly screen out poor vaccine candidates, enable extended evaluation of promising candidates and provide key information on the immunological basis. This talk will provide various perspectives in utilizing adaptive trial design methods for use in vaccine clinical trials; discuss adaptive designs from regulatory prospective including the recommendations from the FDA guidance. The speaker will provide examples of adaptive designs used in vaccine clinical trials and describe the strength and weaknesses of this new approach from statistical and regulatory perspectives in his experience as statistical reviewer of vaccines at the FDA.

email: ghideon.ghebregiorgis@fda.hhs.gov

DETERMINING WHICH SUBPOPULATIONS BENEFIT FROM A VACCINE, USING ADAPTIVE DESIGNS

Michael Rosenblum, Johns Hopkins Bloomberg School of Public Health*

It is a challenge to evaluate experimental treatments, and in particular vaccines, where it is suspected that the treatment effect may only be strong for certain subpopulations, such as those in a certain age range or those having certain risk factors. Standard randomized controlled trials can have low power in such situations. They also are not optimized to distinguish which subpopulations benefit from a treatment. With the goal of overcoming these limitations, we consider randomized trial designs in which the criteria for patient enrollment may be changed, in a preplanned manner, based on interim analyses. Since such designs allow data-dependent changes to the population enrolled, care must be taken to ensure strong control of the familywise Type I error rate, to minimize bias, and to ensure results are clearly interpretable. We present a method for hypothesis testing, point estimation, and confidence intervals tailored for optimizing the information learned about overall and subpopulation treatment effects.

email: mrosenbl@jhsph.edu

64. MIXING: INFERENCES USING FREQUENTIST AND BAYESIAN METHODS AND FOR MIXED DISCRETE AND CONTINUOUS DATA

FLEXIBLE RANDOM EFFECTS COPULA MODELS FOR CLUSTERED MIXED OUTCOMES: APPLICATION IN DEVELOPMENTAL TOXICOLOGY

Alexander R. de Leon, University of Calgary*

The talk concerns the analysis of clustered data with mixed bivariate responses, i.e., where each member of the cluster has a discrete and a continuous outcome. A copula-based random effects model is proposed that accounts for associations between discrete and/or continuous outcomes within and between clusters, including the intrinsic association between the mixed outcomes for the same subject. The approach yields regression parameters in models for both outcomes that are marginally meaningful; in addition, by assuming a latent variable framework to describe discrete outcomes, complications that arise from direct applications of copulas to discrete variables are avoided. Maximum likelihood estimation of the model parameters is implemented using readily available software (e.g., PROC NLMIXED in SAS), and results of simulations concerning the bias and efficiency of the estimates are reported. The proposed methodology is motivated by and illustrated using a developmental toxicity study of ethylene glycol (EG) in mice.

email: adeleon@math.ucalgary.ca

JOINT ANALYSIS OF BINARY AND QUANTITATIVE TRAITS WITH DATA SHARING AND OUTCOME-DEPENDENT SAMPLING

Jungnam Joo, National Cancer Center, Korea*

We study the joint analysis for testing the association between a genetic marker with both binary and quantitative traits, where the quantitative trait values are only available for the cases due to data sharing and outcome-dependent sampling. Data sharing becomes common in genetic association studies, and the outcome-dependent sampling is the consequence of data sharing under which a phenotype of interest is not measured for some subgroup. The trend test and F test are often respectively used to analyze the binary and quantitative traits. Due to the outcome-dependent sampling, the usual F test can be applied using the subgroup with the observed quantitative traits. We propose a modified F test by also incorporating the genotype frequencies of the subgroup whose traits are not observed. Further, a combination of this modified F test and Pearson's test is proposed by Fisher's combination of their p-values as a joint analysis. Due to the correlation of the two analyses, we propose to use a Gamma distribution to fit the asymptotic null distribution for the joint analysis. The proposed modified F-test and the joint analysis can also be applied to test single trait association.

email: jungnam.joo@gmail.com

BAYES FACTOR BASED ON A MAXIMUM STATISTIC FOR CASE-CONTROL GENETIC ASSOCIATION STUDIES

Linglu Wang, The George Washington University*

We study a hybrid Bayes factor for case-control genetic association studies. The proposed Bayes factor models the asymptotic distributions of a robust test statistic under the null and alternative hypotheses. The robust test considered here is the maximum of three trend tests derived under the recessive, additive and dominant genetic models, respectively, referred to as MAX3. To calculate the Bayes factor of MAX3, the asymptotic distribution of MAX3 under the alternative hypothesis is derived. Our proposed Bayes factor is compared to the Bayesian model averaging (BMA). Through simulation studies, we show that our proposed Bayes factor and the BMA are both robust to the unknown genetic model. Although both the proposed method and the BMA depend on the prior for the genetic model, the proposed method is more robust to the choice of the prior than the BMA. Applications to real data are present to illustrate the use of the proposed Bayes factor and also demonstrate that Bayes factor is a better measure than p-value when one comparies results across genetic studies.

email: linglu@gwmail.gwu.edu

ANALYSIS OF CASE-CONTROL QUALITATIVE AND QUANTITATIVE TRAIT DATA FOR GENETIC ASSOCIATION

Minjung Kwak, National Heart Lung and Blood Institute, National Institutes of Health*

We consider a linear model for a quantitative trait in genetic association case-control study. Y is the quantitative trait which is observed only for cases and correlated with case-control status, e.g. biomarker measurement. Y may be probably related to disease status but we are not sure if Y is the causal determinant of case-control status. We impute Y of controls with a fixed y* and fit a linear model postulating that Y of controls tends to be smaller than Y of cases. We set up a likelihood function and estimate the effect of genotype possibly with other covariates in the model. We also discuss how to choose y* and its impact in numerical studies.

email: kwakm2@nhlbi.nih.gov

HYBRID INFERENCE FOR ASSOCIATION STUDIES

Qizhai Li, Academy of Mathematics and Systems Science, Chinese Academy of Sciences*
Jing Qin, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Ao Yuan, Howard University

We consider a Bayesian-frequentist hybrid approach for the analysis of casecontrol genetic association studies. For the parameter of interest, the log-odds ratio of the genetic effect, a Bayesian inference is used, while for the other parameters for covariates, a frequentist method is employed. This hybrid approach is appealing compared to a classical frequentist approach because previous association studies of the same genetic marker enable one to obtain an informative prior for the genetic effect, while such information is often not available for the covariates. It is also computationally simpler than a full Bayesian analysis. We consider both hybrid estimates and hybrid hypothesis testing, the latter is based on a hybrid likelihood function. The asymptotic properties of the hybrid inference are derived, which show the hybrid estimates under common loss functions are first-order equivalent to the maximum likelihood estimates and fully efficient, and that the hybrid likelihood ratio test and hybrid score test have the same asymptotic distributions as the original likelihood ratio test and score test under the null hypothesis.

email: liqz@amss.ac.cn

65. BAYESIAN METHODS FOR LONGITUDINAL AND/OR SURVIVAL DATA

POSTERIOR PREDICTIVE MODEL ASSESSMENT FOR INCOMPLETE LONGITUDINAL DATA

Arkendu Chatterjee, University of Florida*
Michael Daniels, University of Florida

We examine the operating characteristics of two approaches to assess model fit for incomplete longitudinal data. The first approach assesses fit based on replicated complete data as advocated in Gelman et al. (2005, Biometrics). The second approach assesses fit based on replicated observed data. Pros and cons of each approach are discussed and simulations and analytical results are presented that compare the power under each approach.

email: achatter@stat.ufl.edu

A NOVEL BAYESIAN APPROACH FOR ANALYZING INTERVAL-CENSORED FAILURE TIME DATA UNDER THE PROPORTIONAL HAZARDS MODEL

Xiaoyan Lin, University of South Carolina
Bo Cai, University of South Carolina*
Lianming Wang, University of South Carolina
Zhigang Zhang, Memorial Sloan-Kettering Cancer Center

The proportional hazards model (PH) is the most widely used semiparametric regression model for analyzing time-to-event data. However, its popularity is limited to right-censored data, for which the partial likelihood is available allowing one to estimate the regression coefficients directly without estimating the baseline hazard function. Analyzing interval-censored data is challenging due to the complexity of the data structure, and existing approaches are usually either too technically involved or too computationally expensive, making them impractical for applied statisticians and study investigators. In this paper, we propose an efficient and easy-to-implement Bayesian approach for analyzing interval-censored data under the PH model. Our approach models the cumulative baseline hazard function with monotone splines and allows one to estimate the regression parameters and spline coefficients simultaneously. The proposed Gibbs sampler relies on a novel data augmentation and does not require imputing unobserved failure times or contain any complicated Metropolis-Hastings steps. The proposed method outperforms many existing approaches as shown in our simulation study and is illustrated in a colon cancer data set from Memorial Sloan-Kettering Cancer Center.

email: bcai@sc.edu

SEMIPARAMETRIC BAYESIAN SURVIVAL ANALYSIS USING MODELS WITH LOG-LINEAR MEDIAN

Jianchang Lin, Florida State University*
Debajyoti Sinha, Florida State University
Stuart Lipsitz, Brigham and Women's Hospital
Adriano Polpo, University of São Paulo

We present two semiparametric survival models with log-linear median regression functions as useful alternative to the popular Cox's (1972) models and linear transformation models (Cheng et al., 1995). Compared to existing semiparametric models, our models have many practical advantages, including the interpretation of regression parameters via median and ability to address heteroscedasticity. We demonstrate that our modeling techniques facilitate the ease of prior elicitation and computation for both parametric and semiparametric analysis of survival data. We illustrate modeling advantages, data analysis and model diagnostics via reanalysis of a small-cell lung cancer study.

email: jlin@stat.fsu.edu

A MODEL-BASED APPROACH TO LIMIT OF DETECTION IN STUDYING ENVIRONMENTAL CHEMICAL EXPOSURES AND TIME TO PREGNANCY

Sungduk Kim, 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
Enrique F. Schisterman, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Neil Perkins, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Germaine M. Buck Louis, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Human exposure to persistent environmental pollutants often results in a range of exposures with a proportion of concentrations below laboratory detection limits. Growing evidence supports that inadequate handling of concentrations that are below the limit of detection (LOD) may bias health effects in relation to chemical exposures. We sought to quantify such bias in day specific probability of conception during the fertile window, and propose a model-based approach to reduce the biases. We assume a multivariate skewed generalized t-distribution constrained by LODs, which realistically represents the underlying shape of the chemical exposures. Correlations in the multivariate distribution provided information across chemicals. A Markov chain Monte Carlo sampling algorithm is developed for implementing the Bayesian computations. The deviance information criterion measure is used for guiding the choice of distributions for chemical exposures with LODs. We applied the proposed method to the data from the Longitudinal Investigation of Fertility and the Environment (LIFE) Study.

email: kims2@mail.nih.gov

A SEMIPARAMETRIC BAYESIAN APPROACH FOR JOINT MODELING OF LONGITUDINAL TRAIT AND EVENT TIME: APPLICATION TO SOYBEAN DATA

Kiranmoy Das, Temple University*

Different traits, though appear to function separately, can actually control an event simultaneously. This fundamental biological principle has motivated researchers not to study the traits separately but to develop a joint model to explain the system more

efficiently. Because of advanced biotechnology, huge amount of genetic information can be obtained in current days. In a statistical sense, dimension reduction becomes one of the major issues in biological researches. In this paper, we combine dimension reduction and multiple testing, two key statistical problems in biomedical research in the context of joint modeling of observed longitudinal trait and the event time in a bayesian semi-parametric framework. A sure independence screening procedure based on the distance correlation (DC) followed by a slightly modified version of Bayesian Lasso are used for the dimension reduction and zero-inflated Dirichlet priors are considered in the joint model. We applied the proposed method for detecting genes having significant effect on the time to get the first flower via the biomass for soybean plants. Extensive simulation studies verify the practical applicability and usefulness of the proposed methodology.

email: kiranmoy.das@temple.edu

HIERARCHICAL BAYESIAN APPROACH FOR ANALYSIS OF LONGITUDINAL COUNT DATA WITH OVERDISPERSION PARAMETERS: A SIMULATION STUDY

Mehreteab F. Aregay, University of Leuven, Belgium*
Geert Molenberghs, I-BioStat Belgium
Ziv Shkedy, Hasselt University, Belgium

In observed count data, the sample variance is often considerably larger/smaller than the sample mean which is known as a problem of overdispersion/underdispersion. This article focuses on hierarchical Bayesian modeling of longitudinal count data. Two different models are considered. The first assumed a Poisson distribution for the count data and includes a subject specific intercept (which assumed to follow a normal distribution) in order to account for subject heterogenity. However, such a model ignores the possible problem of overdispersion/underdispersion. The second model which we considered includes, in addition to the random intercept, a random subject and time dependent parameters (assumed to be gamma distributed) which account for overdispersion/underdispersion. In order to compare the performance of the two models a simulation studies were conducted in which the MSE, bias, relative bias and variance of the posterior means are compared.

email: mehreteabfantahun.aregay@med.kuleuven.be

BAYESIAN MODELING DEPENDENCE IN LONGITUDINAL DATA VIA PARTIAL AUTOCORRELATIONS AND MARGINAL VARIANCES

Yanpin Wang, University of Florida*
Michael Daniels, University of Florida

Many parameters and positive-definiteness are two major obstacles in estimating and modelling a correlation matrix for longitudinal data. In addition, when longitudinal data is incomplete, incorrectly modelling the correlation matrix often results in bias in estimating mean regression parameters. In this paper, we introduce a flexible class of regression models for Fisher's z-transform (link) of the partial autocorrelations. The partial autocorrelations proposed can freely vary in the interval (-1,1) while maintaining positive definiteness of the correlation matrix so the regression parameters in these models will have no constraints. We propose a class of priors for the regression coefficients. We examine the importance of correctly modeling the correlation structure on estimation of longitudinal (mean) trajectories via simulations. The regression approach is illustrated on data from a longitudinal clinical trial.

email: yanpin@ufl.edu

66. COMPLEX STUDY DESIGNS AND BIAS CORRECTIONS

TWO-STAGE DESIGNS FOR ADAPTIVE COMPARATIVE EFFECTIVENESS TRIALS

John A. Kairalla, University of Florida*
Mitchell A. Thomann, University of Iowa
Christopher S. Coffey, University of Iowa
Keith E. Muller, University of Florida

Unlike in standard clinical trials, the concept of a "minimal clinically meaningful effect" has little meaning in clinical comparative effectiveness trials. Rather, any reliable difference between two active treatments is clinically meaningful for the population, assuming roughly equal costs and side effects. Detecting smaller effects requires larger sample sizes; however, a trial may only be powered to detect large to moderate differences. Thus, clinically small, but population important, differences may be missed. We examine a class of two-stage studies that allows for stopping as planned under moderate to large effect size, while allowing study continuation and resizing should the observed effect be smaller than anticipated. Exact theory allows changing the planned effect size in the comparative effectiveness setting. The results enable quick calculations for power, type I error rate, and expected sample size for any combination of sample size, true effect size, observed variance, stopping bound types, and timings. Examples illustrate the value of the recommendations presented.

email: johnkair@ufl.edu

MORE EFFICIENT ESTIMATORS FOR CASE-COHORT STUDIES

SoYoung Kim, University of North Carolina at Chapel Hill*
Jianwen Cai, University of North Carolina at Chapel Hill

Case-cohort study design is generally used to reduce cost in large cohort studies. The case-cohort design consists of a random sample of the entire cohort, named subcohort, and all the subjects with the disease of interest. When several diseases are of interest, several case-cohort studies are usually conducted using the same subcohort. When these case-cohort data are analyzed, the

common practice is to analyze each disease separately ignoring data collected in subjects with the other diseases. This is not efficient use of the data. In this paper, we propose more efficient estimators by using all available information. We consider both joint analysis and separate analysis. We propose an estimation equation approach with a new weight function. We establish that the proposed estimator is consist and asymptotically normally distributed. Simulation studies show that the proposed methods using all available information gain efficiency. For comparing the effect of the exposure on different diseases, tests based on the joint analysis are more powerful than those based on the separate analysis. We apply our proposed method to the data from the Busselton Health Study.

email: kimso@live.unc.edu

ESTIMATING MULTIPLE TREATMENTS EFFECTS USING TWO-PHASE REGRESSION ESTIMATORS

Cindy Yu, Iowa State University
Jason Legg, Amgen Inc.
Bin Liu, Iowa State University*

We propose a semiparametric two-phase regression estimator with a semiparametric generalized propensity score estimator for estimating average treatment effects in the presence of informative first-phase sampling. The proposed estimator is shown to be easily extendable to any number of treatments and does not rely on a pre-specified form of the response or outcome functions. The estimator is shown to asymptotically outperform standard estimators such as the double expansion estimator and eliminate bias found in naive estimators that ignore the first-phase sample design such as inverse propensity weighted estimator. Potential performance gains are demonstrated through a simulation study.

email: lbbb@iastate.edu

A SEMI-PARAMETRIC APPROACH TO SELECT OPTIMAL SAMPLING SCHEDULES FOR MEASURING THE MEAN PROFILE AND VARIABILITY IN LONGITUDINAL STUDIES

Meihua Wu, University of Michigan*
Brisa N. Sánchez, University of Michigan
Trivellore E. Raghunathan, University of Michigan
Ana V. Diez-Roux, University of Michigan

Longitudinal studies are frequently used to investigate the patterns of health outcomes over time. A critical component of longitudinal study design involves determining the sampling schedule. Criteria for optimal design often focus on estimation of the mean profile, but the estimation of variance components of the longitudinal process is also important, since variance patterns may be associated with covariates of interest or predict future outcomes. Existing approaches, based on parametric mixed models (PMM), have limited applicability when one wishes to accurately estimate the mean and variance parameters simultaneously. We propose a semiparametric model to separately characterize the mean profile and the variability, and use it to derive optimal sampling schedules. We use functional principal component analysis (FPCA) to regularize the variability process, leading to a parsimonious and flexible representation of the temporal pattern of the variability. Simulation studies suggest that the new approach outperforms the schedules derived by PMM in certain cases. We employ the new approach in two applications (salivary cortisol and urinary progesterone). Our method is shown to identify sampling regions that are not discovered by the PMM approach and select sampling schedules that are predictive for subsequent outcome.

email: meihuawu@umich.edu

AN IMPROVED PAIRED AVAILABILITY DESIGN FOR HISTORICAL CONTROLS

Stuart G. Baker, National Cancer Institute, National Institutes of Health*
Karen S. Lindeman, Johns Hopkins University

When a randomized trial cannot be implemented for a comparative effectiveness analysis, an appealing alternative in the some applications is the paired availability design for historical controls (Baker and Lindeman, 1994). The goal of the paired availability design is to estimate the effect of receipt of treatment using data from multiple medical centers over two time periods with a change in availability of treatment. Recent work has refined the assumptions and introduced an improved estimate of treatment effect that improves generalizability from the principal strata of interest to all eligible persons. The application involved the effect of epidural analgesia on the probability of Caesarean section. The improved estimate from the paired availability design was similar to the estimate from a meta-analysis of randomized trials and differed considerably from an estimate based on the propensity scores computed in another study.

email: sb16i@nih.gov

BIAS CORRECTION AND LIKELIHOOD BASED INFERENCE UNDER MODEL MISSPECIFICATION

Yang Ning, Johns Hopkins University*
Kung-Yee Liang, National Yang-Ming University

How to correct the potential bias of the estimator under model misspecification is an important topic in statistics. In reality, the true model which characterizes the underlying phenomenon may be complicated. Due to the heavy computational burden or potential misspecification of the true model, a simpler but misspecified model is often used in practice. In the current paper, we propose a unified approach to correct the bias of the estimator and perform the likelihood inference based on the misspecified

model as long as some information about the true model is available. In particular, to deal with the nuisance parameters, we introduce a pseudo likelihood approach as well as a sensitivity analysis approach. The corresponding asymptotic properties are examined and the finite sample performance is considered through several examples, simulations and real data analysis.

email: yning@jhsph.edu

67. HIGH DIMENSIONAL DATA

SPARSE META-ANALYSIS WITH APPLICATIONS TO HIGH-DIMENSIONAL DATA

Qianchuan He, University of North Carolina at Chapel Hill*
Helen Hao Zhang, North Carolina State University
Danyu Lin, University of North Carolina at Chapel Hill
Christy L. Avery, University of North Carolina at Chapel Hill

Meta-analysis plays an important role in summarizing and synthesizing scientific evidence from multiple studies. When the dimensions of the data are high, it is desirable to incorporate variable selection intometa-analysis to improve model interpretation and prediction. Existing variable selection methods require direct access to raw data, and many of them assume the effects of a feature to be the same across studies. We propose a novel method, Sparse Meta-Analysis (SMA), that is able to conduct variable selection for meta-analysis solely based on summary statistics. In addition, our method allows the effect sizes of a feature to vary among studies. We show that our method enjoys selection consistency and the oracle property as if the raw data were available. Simulations and real data analysis demonstrate that SMA performs well in both variable selection and prediction. Since summary statistics are far more accessible than raw data, our method has broader applications in high-dimensional meta-analysis.

email: heqianch@email.unc.edu

UNIVERSAL PROBABILISTIC DEPENDENCY DISCOVERY: THEORY AND APPLICATION

Hesen Peng, Emory University*
Yun Bai, Philadelphia College of Osteopathic Medicine
Tianwei Yu, Emory University

The emergence of high-throughput data in biological science and computer networks has generated novel challenges for statistical methods. Nonlinear relationships and dependencies involving multiple variables are abundant. The sheer volume of high-throughput data have limited the application for traditional case-by-case analysis methods, whose model assumptions, like linearity, are not supported in high-throughput scenarios. To meet these challenges, we developed Mira score, a novel probabilistic dependency measure that accounts for probabilistic dependency of arbitrary dimension and arbitrary type. Mira score is defined as a function of observation graph, and thus circumvents the curse of dimensionality in high-dimensional data. The superior statistical property enjoyed by Mira score has lead to our development of efficient network reverse-engineering procedure for multivariate dependencies. As an example, the procedure has been applied to celiac disease, and lung cancer pathway interaction analysis. The study found the interaction between ATP-binding cassette transporters and nitrogen metabolism pathways suppressed in celiac disease patients. For the lung cancer case, our analysis found the interaction between nicotinate and nicotinamide metabolism and caffeine metabolism pathways amplified in lung cancer patients.

email: hesen.peng@gmail.com

INVESTIGATING PYROSEQUENCE DATA FROM ECOLOGICAL APPLICATIONS

Karen Keating, Kansas State University*
Gary L. Gadbury, Kansas State University
Ari Jumpponen, Kansas State University
Karen A. Garrett, Kansas State University

Since their commercial introduction in 2005, DNA pyrosequencing technologies have become widely available and are now cost-effective tools for determining the genetic characteristics of organisms. While the biomedical applications of DNA sequencing are apparent, these technologies have been applied to many other research areas. One such area is community ecology, in which DNA sequence data are used to identify the presence and abundance of microscopic organisms that inhabit an environment. This is currently an active area of research, since it is generally believed that a change in the composition of microscopic species in a geographic area may signal a change in the overall health of the environment. We present an overview of DNA sequencing-by-synthesis as implemented by the Roche/Life Science 454 platform, and identify aspects of this process that can introduce variability in the data. We also examine four ecological data sets generated by the 454 platform, with particular emphasis on the unique characteristics of these data, and explore methods for identifying and mitigating excessive variation in the data.

email: keatingk@ksu.edu

EXPLORATION OF REACTANT-PRODUCT LIPID PAIRS IN MUTANT-WILD TYPE LIPIDOMICS EXPERIMENTS

Lianqing Zheng, Kansas State University*
Gary L. Gadbury, Kansas State University
Jyoti Shah, University of North Texas
Ruth Welti, Kansas State University

As "omics" high-throughput metabolite profiling is developed, developing methodology to use the data to identify the functions of genes is very important to biologists. For genes that encode enzymes, a mutation in the gene is expected to alter the level of the metabolites which serve as the enzyme's reactant(s) (also known as substrate) and product(s). Using metabolite data from a wild-type organism and one with a gene silenced by a mutation, the goal is to identify candidate metabolites for the normal reactants and products of the enzyme that was genetically altered. Comparing a mutant organism to a wild-type organism, the reactant concentration level will be higher and the product concentration level lower in the mutant. This is because the effect of the mutation is to block the reaction between reactant and product. To detect possible reactant and product pairs in a lipidomics experiment done in a plant (Arabidopsis thaliana) system, based on the above scheme, we have developed a technique using several test statistics. Parametric null distributions of the test statistics are derived using a bootstrap method to obtain distributional characteristics of the test statistics under a null hypothesis. This then forms the basis for a test for detecting product-reactant pairs in a mutant-wild type lipdomics experiment.

email: lzheng@ksu.edu

FACTOR ANALYSIS REGRESSION FOR PREDICTIVE MODELING WITH HIGH DIMENSIONAL DATA

Netsanet T. Imam, State University of New York at Buffalo*
Randy L. Carter, State University of New York at Buffalo
Russell W. Bessette, State University of New York at Buffalo

We present factor-model based method to predict a response, y, as a linear function of explanatory variables, x = (x1 ,x2,...,xp) , where the sample size, n, is less than p. We estimated the coefficient parameters of the model using bivariate factor analysis. A Monte Carlo (MC) study was performed to compare our factor analysis (FA) regression with partial least squares (PLS) regression under three underlying data structures: arbitrary correlation, factor model correlation structure, and when y is independent of x. Under each structure, we generated 500 MC training samples, for each sample from a multivariate normal distribution where the unspecified parameters of each structure were fixed at estimates obtained from analysis of a real dataset, assuming the parameter restrictions of the respective structure. Given the independence structure, we observed sever over-fitting by PLS regression compared to FA regression. In the two cases where there was a relationship between y and x, FA regression has slightly better average mean square error of prediction than PLS regression, mainly when n is very small.

email: ntimam@buffalo.edu

68. HIGH DIMENSIONAL DATA: MACHINE LEARNING, MULTIVARIATE METHODS AND COMPUTATIONAL METHODS

MAJORIZATION MINIMIZATION BY COORDINATE DESCENT FOR CONCAVE PENALIZED GENERALIZED LINEAR MODELS

Dingfeng Jiang, University of Iowa*
Jian Huang, University of Iowa

Recent studies have demonstrated theoretical attractiveness of a class of concave penalties in variable selection, including the smoothly clipped absolute deviation and minimax concave penalties. The computation of concave penalized solutions, however, is a difficult task. We propose a majorization minimization by coordinate descent (MMCD) algorithm for the computation of concave penalized solutions in generalized linear models. In contrast to the existing algorithms that use local quadratic or local linear approximation for the penalty function, the MMCD algorithm seeks to majorize the negative log-likelihood by a quadratic loss, but does not use any approximation to the penalty. This strategy makes it possible to avoid the computation of an scaling factor in each update of the estimates, which improves the efficiency of coordinate descent approach. Under certain regularity conditions, we establish the theoretical convergence properties of the MMCD algorithm. We implement the MMCD algorithm for a penalized logistic regression model using the SCAD and MCP penalties. Simulation studies and a data example indicate that the MMCD algorithm works sufficiently fast for the penalized logistic regression in high-dimensional settings where the number of covariates is much larger than the sample size.

email: dingfeng-jiang@uiowa.edu

REDUCING DIMENSION TO IMPROVE COMPUTATIONAL EFFICIENCY IN HIGH DIMENSIONAL STUDIES

Kevin K. Dobbin, University of Georgia*

The computational overhead of statistical procedures in high dimensions is often very high, discouraging the adoption and use of best statistical practices in data analysis. The high computational cost comes from calculations and manipulations associated with high dimensional vectors and matrices that Monte Carlo, bootstrap and permutation procedures require. We present methods for modeling these high dimensional procedures in a lower dimensional space; then computations can be carried out entirely in low dimensions. We show that the computational savings that results from this approach can be several orders of magnitude, converting an analysis that previously required days to one that requires seconds to execute. We discuss challenges and potential future directions for this research.

email: dobbinke@uga.edu

ADDITIVE KERNEL MACHINE REGRESSION BASED ANALYSIS OF GENOMIC DATA

Jennifer Clark, University of North Carolina at Chapel Hill*
Mike Wu, University of North Carolina at Chapel Hill

High throughput biotechnology has led to a revolution within modern biomedical research. New studies offer researchers an intimate understanding of how genetic features influence disease outcomes and hold the potential to comprehensively address key biological, medical, and public health questions. However, the high-dimensionality of the data, limited availability of samples, and poor understanding of how genomic features influence outcomes are a challenge for statisticians. The field needs powerful new statistical methods to accommodate complex, high dimensional data. Multi-feature testing, where related features are grouped and the cumulative effect tested, is a useful strategy for genomic analyses. Existing methods generally require adjusting for covariates (genomic or environmental factors) in a linear fashion. We propose to model the features and the complex covariates using the flexible additive kernel machine regression (AKMR) framework. We establish and exploit a connection with linear mixed models to allow for estimation and testing within AKMR. We demonstrate that our approach allows for accurate modeling and, improved power to detect true multi-feature effects.

email: jjclark@live.unc.edu

VARIABLE SELECTION FOR HIGH-DIMENSIONAL MULTIVARIATE OUTCOMES WITH APPLICATION TO GENETIC PATHWAY/ NETWORK ANALYSIS

Tamar Sofer, Harvard School of Public Health*
Lee Dicker, Rutgers University
Xihong Lin, Harvard School of Public Health

We consider variable selection for high-dimensional multivariate regression using penalized likelihoods when the number of outcomes and the number of covariates might be large. To account for within-subject correlation, we consider variable selection when a working precision matrix is used and when the precision matrix is jointly estimated using a two-stage procedure. We show that under suitable regularity conditions, penalized regression coefficient estimators are consistent for model selection for an arbitrary working precision matrix, and have the oracle properties and are efficient when the true precision matrix is used or when it is consistently estimated using sparse regression. We develop an efficient computation procedure for estimating regression coefficients using the coordinate descent algorithm in conjunction with sparse precision matrix estimation using the graphical LASSO (GLASSO) algorithm. We develop the Bayesian Information Criterion (BIC) for estimating the tuning parameter and show that BIC is consistent for model selection. We evaluate finite sample performance for the proposed method using simulation studies and illustrate its application using the type II diabetes gene expression pathway data.

email: tsofer@hsph.harvard.edu

ENHANCEMENTS OF SPARSE CLUSTERING WITH RESAMPLING

Wenzhu Bi, University of Pittsburgh*
George C. Tseng, University of Pittsburgh
Julie C. Price, University of Pittsburgh
Lisa A. Weissfeld, University of Pittsburgh

Datasets where p, the number of variables, is significantly larger than the sample size, n, are commonly generated in a large variety of scientific disciplines. These datasets result from the difficulties of subject recruitment and/or the financial burden of the actual data collection in fields such as imaging and genetic analysis. Since many of the datasets of interest in genetics and imaging arise from the development of new technologies, there is little knowledge of groups and/or subsets of subjects within the population that may be of interest, requiring the use of unsupervised learning techniques such as clustering methods. Clustering has seen wide use in the areas of genetics leading to improvements in the methodology in recent years. We propose a method to add resampling onto sparse clustering to improve upon the current clustering methodology. The addition of resampling methods to sparse clustering results in variable selection that is more accurate. The method is also used to assign an "observed proportion of cluster membership" to each observation, providing a new metric by which to measure membership certainty. The performance of the method is studied via simulation and illustrated in the motivating data example.

email: web10@pitt.edu

GENERALIZED REDUCED RANK REGRESSION FOR MULTIVARIATE RESPONSE

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

The common approaches for dimension reduction in high-dimensional data include variable selection and penalized regression. Penalized methods received much attention in recent years, largely due to the flood of high-dimensional data. Approaches like lasso, adaptive lasso, SCAD, and Bayesian lasso has been developed for both the mean parameters and covariance parameters. A less explored approach for multivariate response involves dimension reduction via reduced rank decomposition of the regression coefficient matrix to take advantage of two-way correlations among the regression coefficients. We first derive the framework for L_1 priors on multivariate coefficient matrix in traditional approach. Then we develop the Generalized Reduced rank Regression (GRR) model under L_2 priors and derive the framework for L_1 priors. Simulations and application to ADNI data suggest that GRR has great advantage over traditional approaches; it greatly reduces the number of parameters while performing much better with better comparative advantage for higher dimensions.

email: zak.khondker@parexel.com

69. VARIABLE AND MODEL SELECTION METHODS

SIMULTANEOUS RANK DETERMINATION AND VARIABLE SELECTION IN MULTIVARIATE REDUCED-RANK REGRESSION

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

Chen et al. (2011) developed a novel approach for regularized reduced-rank regression with sparse singular value decomposition (RRR-SSVD). A key attraction of the RRR-SSVD method is that it performs variable selection of both the multivariate response and multivariate predictor while preserving the reduced-rank structure. However, the method requires the strong assumption of known rank. Here we extend the RRR-SSVD approach to conduct simultaneous rank determination and sparse SVD estimation. We show that under mild regularity conditions, both the rank and the sparse SVD structure of the coefficient matrix can be correctly identified with probability approaching one as sample size increases. The empirical performance of the method is investigated via simulation studies. We analyze a macro-economical time series data set to demonstrate the efficacy of the proposed method in reduced-rank vector autoregressive (VAR) modelling.

email: kunchen@ksu.edu

VARIABLE SELECTION FOR FIXED AND RANDOM EFFECTS IN MULTILEVEL MODELS WHEN MISSING DATA IS PRESENT

Miguel Marino, Harvard University*
Yi Li, University of Michigan

Typical variable selection procedures require complete data to be observed (i.e. no missing data). How does one perform variable selection when missing data is present? Practically, this question arises from our work with a cancer prevention study that seeks to identify relevant predictors associated with fruit and vegetable consumption. We develop a procedure that is able to perform fixed and random effects selection for multilevel models with missing data in the covariates. We address the missing data issue with a multiple imputation approach. The presence of multiple datasets creates additional challenges of combining variable selection results across multiple data sets. We propose a novel approach that stacks the multiply-imputed data sets which can allow the use of group variable selection via the group lasso to assess the overall significance of each predictor across the imputed data sets.

email: mmmiguelmm@gmail.com

VARIABLE SELECTION IN PARAMETRIC AND NON-PARAMETRIC REGRESSION

Trang T. Duong, The University of West Georgia*

Variable selection is critical to high-dimensional statistic modeling. Many approaches are used for variable selection such as LASSO (Tibshirani, 1996), SCAD (Fan and Li 2001) and MCP (Zhang 2010). However, for low dimensional models, the least square method is more popular. In our project, we apply such penalized methods for high-dimensional cases to low dimensional cases. We compare different penalized methods for both parametric models and nonparametric models in terms of residual sum of squares, the number of parameters selected, correct selection percentage, over selection percentage and lower selection percentage. From our simulation studies, we can see that SCAD and MCP have the oracle selection property that was well understood. The result of LASSO although is better than the least square method but the accuracy of the model given by LASSO is not good as ones given by SCAD and MCP. Moreover, we apply such methods to solve some real problems.

email: tduong1@my.westga.edu

PENALIZED VARIABLE SELECTION WITH U-ESTIMATES

Xiao Song, University of Georgia*
Shuangge Ma, Yale University School of Public Health

U-estimates are defined as maximizers of objective functions that are U-statistics. As an alternative to M-estimates, U-estimates have been extensively used in linear regression, classification, survival analysis, and many other areas. They may rely on weaker data and model assumptions and be preferred over alternatives. In this article, we investigate penalized variable selection with U-estimates. We first propose smooth approximations of the objective functions, which can greatly reduce the computational cost without affecting the asymptotic properties. Instead of attempting to create any new penalties, we focus on penalized variable selection with U-estimates using penalties that have been well investigated with M-estimates, including the LASSO, adaptive LASSO and bridge, and establish their asymptotic properties. Generically applicable computational algorithms are described. Performance of the penalized U-estimates is assessed using numerical studies.

email: xsong@uga.edu

VARIABLE SELECTION WITH ITERATED PENALIZATION FOR SEMIPARAMETRIC REGRESSION MODELS

Ying Dai, Yale University School of Public Health*
Shuangge Ma, Yale University School of Public Health

Consider semiparametric regression analysis with a moderate to large number of covariates. When there are covariates that may be not associated with the response variable, variable selection is needed along with model estimation. In this study, we adopt a sieve approach with a diverging number of basis functions for estimation of the nonparametric covariate effects in semiparametric regression models. We propose an iterated penalization approach for regularized estimation and variable selection. In the first step, a mixture of Lasso and group Lasso penalties are employed to obtain the initial estimate. In the second step, a mixture of weighted Lasso and weighted group Lasso penalties, where the weights are constructed using the initial estimate, are employed. We show that the proposed iterated approach can lead to consistent variable selection, even when the number of unknown parameters diverges at a rate faster than the sample size. Numerical study, including simulation and analysis of a diabetes dataset, shows satisfactory finite-sample performance of the proposed approach.

email: ying.dai@yale.edu

SPARSITY RECOVERY FROM MULTIVARIATE SMOOTHING FUNCTIONS USING THE NONNEGATIVE GARROTE METHOD

Zaili Fang, Virginia Polytechnic Institute and State University*
Inyoung Kim, Virginia Polytechnic Institute and State University
Patrick Schaumont, Virginia Polytechnic Institute and State University

We propose a nonnegative garrote on kernel machines (NGK) method to recover the sparsity of input variables in a multivariate smoothing function. We model the smoothing function by a least squares error kernel machine. Since kernel matrix can be expressed as a componentwise function of the multivariate similarity matrix which can be written as an additive form of univariate similarity matrix in terms of p input variables, and adding a sparse scale parameter on each input variable indicates whether that variable is relevant to the response, we can apply nonnegative garrote constraint on those scale parameters to construct the NGK model. With different kernel structures, our method can be applied to either additive or nonadditive models. In theoretical aspects, we analyze the asymptotic properties of NGK, and conclude it is a square root consistent estimator of the scale parameters. We further analyze the sufficient and necessary conditions for the sparsistency of NGK given the true initial kernel function coefficients, and we show the sparsistency is satisfied with consistent initial kernel function coefficients under certain conditions. An efficient coordinate descent/backfitting algorithm is introduced.

email: zlfang@vt.edu

70. PRESIDENTIAL INVITED ADDRESS

ENGAGING, INSPIRING, AND TRAINING THE NEXT GENERATION: PAST SUCCESSES, FUTURE CHALLENGES AND OPPORTUNITIES

Marie Davidian, North Carolina State University

Our discipline is in an unprecedented and enviable position. Scientific inquiry, public policy, and decision-making in industry are all increasingly dependent on the collection and interpretation of what are often vast amounts of complex data, and we – statisticians – are uniquely qualified to address the challenges posed by this data "explosion" and to ensure that the inferences drawn are sound and that the results are communicated appropriately. Opportunities for statisticians abound; the position of statistician has even been called "the sexy job in the next ten years." Advanced Placement (AP) statistics courses in high school have seen a tremendous rise in enrollment in the past decade. So why aren't more US students pursuing graduate training in our discipline and choosing statistics as a career? My experience and that of numerous colleagues in academia, industry, and government is that many qualified US students still do not know enough about the opportunities for statisticians or the training required and are diverted by other Science, Technology, Engineering, and Mathematics (STEM) disciplines that may be more familiar.

This shortage of US students entering our graduate programs and profession is nothing new. For example, two workshops were held by NIH in the early 2000s to discuss the need for increasing the pipeline of biostatisticians to meet the expanding needs of the nation's health sciences research enterprise and resulted in a white paper (DeMets et al. 2006) calling for more training programs and opportunities to encourage US students to pursue biostatistics careers. In 2003, the National Heart, Lung, and Blood Institute (NHLBI) took action, soliciting applications for a "Summer Institute for Training in Biostatistics" (SIBS), restricted to US citizen and permanent resident undergraduates, to expose these students to biostatistical science and practice and the myriad career opportunities available and to encourage them to seek graduate training. What began as three such programs in 2004 was expanded to eight in 2010, and over the past eight summers, hundreds of students have participated, and scores who might otherwise have pursued training in other STEM disciplines have entered graduate programs in statistics and biostatistics nationwide. However, this and the small number of other government-funded statistics programs cannot alone address the challenge we face in bringing talented, diverse students to our field.

Since 2004, I have been privileged to co-direct one of the eight SIBS programs, which is a joint effort between my Department and Duke Clinical Research Institute (DCRI). I also direct a NHLBI-funded predoctoral training program that provides US

PhD students in my department with unparalleled collaborative experience at DCRI. I have seen firsthand how such opportunities have been transformative, altering the career aspirations of so many US students. In this talk, I will review the history of all eight SIBS programs and my experience with training the next generation more generally. I will then argue that, if we are to achieve the statistical workforce required to meet the demand, there must be a broader effort in which stakeholders from all sectors, industry, government, and academia, come together to conceive of and support programs to increase the numbers of US students entering graduate programs in statistics and biostatistics and to provide them with essential practical experience and skills while they are still in training. I hope to inspire all of you to join me in making such an effort a reality.

DeMets, D.L., Stormo, G., Boehnke, M., Louis, T.A., Taylor, J., and Dixon, D. (2006). Training of the next generation of biostatisticians: A call to action in the U.S. Statistics in Medicine 25, 3415–3429.

e-mail: marie_davidian@ncsu.edu

71. RECENT ADVANCES IN STATISTICAL METHODS FOR DIAGNOSTIC MEDICINE

SEMIPARAMETRIC ESTIMATION OF THE COVARIATE-SPECIFIC ROC CURVE IN PRESENCE OF IGNORABLE AND NON-IGNORABLE VERIFICATION BIAS

Xiao-Hua Andrew Zhou, University of Washington*
Danping Liu, University of Washington

In estimation of the ROC curve, when the true disease status is subject to nonignorable missingness, the observed likelihood involves the missing mechanism given by a selection model. In this presentation, we describe several new semi-parametric methods to estimate the ROC curve and the area under ROC curve when the verification bias is nonignorable. We have also proposed a semiparametric method for estimating the covariate-specific ROC curves with a partial missing gold standard. Three new ROC curve estimators are proposed and compared, namely, imputation-based, inverse probability weighted and doubly robust estimators. We derive the asymptotic normality of the estimated ROC curve, as well as the analytical form the standard error estimator. The proposed method is motivated and applied to the data in an Alzheimer's disease research.

email: azhou@uw.edu

ESTIMATION AND DESIGN FOR LOGISTIC REGRESSION UNDER AN IMPERFECT POPULATION IDENTIFIER

Paul S. Albert, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*
Aiyi Liu, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Tonia Nansel, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

We consider estimation and design for logistic regression when the population is identifed with a test which is measured with error. We propose a maximum-likelihood approach for parameter estimation when the imperfect test is measured on all study participants and the gold standard test is only observed on a small subset of individuals. Under maximum-likelihood estimation, we evaluate the optimal design in terms of sample selection as well as verification. We show that there may be substantial efficiency gains by choosing a small percentage of individuals who test negative on the imiperfect test for inclusion in the sample. Further, we show that there is little efficiency gain for verifying test postives versus test negatives with differing probabilities. Alternative approaches including mean imputation, re-weighted, and semi-parametric efficient estimation are also considered. We compare the estmation approaches under different designs with simulations. The methodology is illustrated with an analysis from a diabetes behavioral intervention trial.

email: albertp@mail.nih.gov

DESIGNING STUDIES TO EVALUATE BIOMARKERS FOR SELECTING PATIENT TREATMENT

Holly Janes, Fred Hutchinson Cancer Research Center*
Margaret Pepe, Fred Hutchinson Cancer Research Center
Ying Huang, Fred Hutchinson Cancer Research Center
Marshall Brown, Fred Hutchinson Cancer Research Center

Biomarkers associated with patient response to treatment have the potential to improve clinical outcomes by restricting treatments to the patients most likely to benefit. The ideal setting for evaluating a treatment selection biomarker is a randomized controlled trial. The biomarker may be measured at baseline on the entire trial population, or on a subset of participants potentially selected on the basis of treatment response. Existing study design methodology is limited, and focuses on evaluating biomarkers by testing for a statistical interaction between biomarker and treatment assignment. We propose an approach which sizes the study to evaluate the impact of a marker-based treatment policy on the population response rate. We provide methods for determining the required number of patients and optimal treatment allocation and for sub-sampling from the trial population based on treatment response. We illustrate the approach using a study to evaluate the Oncotype DX marker for selecting adjuvant chemotherapy to treat estrogen-receptor positive breast cancer.

email: hjanes@scharp.org

72. JABES SPECIAL SESSION ON CLIMATE CHANGE AND HEALTH

ESTIMATING THE HEALTH IMPACT OF CLIMATE CHANGE WITH CALIBRATE CLIMATE MODEL OUTPUT

Montse Fuentes, North Carolina State University*

Studies on the health impacts of climate change routinely use climate model output as future exposure projection. Uncertainty quantification, usually in the form of sensitivity analysis, has focused predominantly on the variability arise from different emission scenarios or multi-model ensembles. This paper describes a Bayesian spatial quantile regression approach to calibrate climate model output for examining to the risks of future temperature on adverse health outcomes. Specifically, we first estimate the spatial quantile process for climate model output using nonlinear monotonic regression during a historical period. The quantile process is then calibrated using the quantile functions estimated from the observed monitoring data. Our model also down-scales the gridded climate model output to the point-level for projecting future exposure over a specific geographical region. The quantile regression approach is motivated by the need to better characterize the tails of future temperature distribution where the greatest health impacts are likely to occur. We applied the methodology to calibrate temperature projections from a regional climate model for the period 2041 to 2050. Accounting for calibration uncertainty, we calculated the number of of excess deaths attributed to future temperature for three cities in the US state of Alabama.

email: fuentes@ncsu.edu

FLEXIBLE DISTRIBUTED LAG MODELS USING RANDOM FUNCTIONS WITH APPLICATION TO ESTIMATING MORTALITY DISPLACEMENT FROM HEAT-RELATED DEATHS

Roger Peng, Johns Hopkins University*

Changes in the distribution of ambient temperature, due to climate change or otherwise, will likely have a negative effect on public health. Characterizing the relationship between temperature and mortality is a key aspect of the larger problem of understanding the health effect of climate change. In this article, a flexible class of distributed lag models are used to analyze the effects of heat on mortality in four major metropolitan areas in the U.S. (Chicago, Dallas, Los Angeles, and New York). Specifically, the proposed methodology uses Gaussian processes to construct a prior model for the distributed lag function. Gaussian processes are adequately flexible to capture a wide variety of distributed lag functions while ensuring smoothness properties of process realizations. The proposed framework also allows for probabilistic inference of the maximum lag. Applying the proposed methodology revealed that mortality displacement (or, harvesting) was present for most age groups and cities analyzed suggesting that heat advanced death in some individuals. Additionally, the estimated shape of the DL functions gave evidence that highly variable temperatures pose a threat to public health.

email: rpeng@jhsph.edu

A COMPARTMENTAL MODEL FOR MENINGITIS: SEPARATING TRANSMISSION FROM CLIMATE EFFECTS ON DISEASE

Roman Jondarov, Penn State University*
Murali Haran, Penn State University
Matthew Ferrari, Penn State University

Every year countries of the African meningitis belt are afflicted with meningococcal meningitis disease outbreaks. The timing of these outbreaks coincide with the dry season. There are two main hypotheses about this strong seasonal effect. The first hypothesis assumes that there is a seasonally forced increase in the risk of transition from being an asymptomatic carrier to an individual with invasive disease. The second hypothesis states that the incidence of meningitis increases due to higher transmission of the infection during the dry season. In this paper, we develop a statistical model to investigate these hypotheses. Standard maximum likelihood or Bayesian inference for this model is infeasible as there are potentially tens of thousands of latent variables in the model and each evaluation of the likelihood is expensive. We therefore propose an approximate Bayesian computation (ABC) based approach to infer the unknown parameters of the model. Our approach allows us to study the marginal and joint posterior distributions of these parameters, thereby allowing us to answer scientific questions of interest. We
apply our modeling and inferential approach to data on cases of meningitis for 34 communities in Nigeria from Médecins Sans Frontières (MSF) and World Health Organization (WHO) for 2009.

email: raj153@psu.edu

BIVARIATE DOWNSCALING WITH ASYNCHRONOOUS MEASUREMENTS

Yunwen Yang, Drexel University*
Xuming He, University of Michigan

Statistical downscaling is a useful technique to localize global or regional climate model projections to assess the potential impact of climate changes. It requires quantifyinga relationship between climate model output and local observations from the past, but the two sets of measurements are not necessarily taken simultaneously, so the usual regression techniquesare not applicable. In the case of univariate downscaling, the Statistical Asynchronous Regression (SAR) method of O'Brien, Sornette and McPherron (2001) provides a simple quantile-matching approach with asynchronous measurements. In this paper, we propose a

bivariate downscaling method for asynchronous measurements based on a notion of bivariate ranks and positions. The proposed method is preferable to univariate downscaling, because it is able to preserve general forms of association between two variables, such as temperature and precipitation, in statistical downscaling. This desirable property of the bivariate downscaling method is demonstrated through applications to simulated and real data.

email: yy365@drexel.edu

73. GRANT FUNDING OPPORTUNITIES FOR BIOSTATISTICIANS

NEW OPPORTUNITIES FOR RESEARCH FUNDING AT NSF

Haiyan Cai, National Science Foundation*

NSF provides many new funding oppotunities for researchers like BioMaPS or Research at the Interface of the Biological, Mathematical and Physical Sciences and Engineering. The novelty of the BioMAPS approach is the strategic investigation of living systems across scales from atoms and molecules to organisms to environment, and the application of that knowledge to develop new fundamental understanding and new technologies. While the topics are not new, recent advances in genomics, synthetic biology, nanotechnology, analytical instrumentation, and computational and data-intensive science and engineering enable us to make significant progress in ways that were not possible even a few years ago. The budget request for BioMaPS in fiscal 2012 is $76 million. Another important new initiative is SAVI, Science Across Virtual Institutes.

email: hcai@nsf.gov

OVERVIEW OF NIH APPLICATION PROCESSES

Michelle C. Dunn, National Cancer Institute, National Institutes of Health*

Michelle Dunn, a Program Officer at the National Cancer Institute, one of 27 institutes that comprise the National Institutes of Health (NIH), will give an overview of NIH application process.
email: dunnm3@mail.nih.gov

PEER REVIEW AT THE NATIONAL INSTITUTES OF HEALTH

Tomas Drgon, Center for Scientific Review, National Institutes of Health*

The Center for Scientific Review (CSR) is responsible for the review of the scientific merit of NIH grant applications. The CSR organizes peer review groups or study sections that evaluate majority of research grant applications sent to NIH. The purpose of the CSR is to see that these applications receive fair, independent, expert, and timely reviews, free from inappropriate influences, so that NIH can fund the most promising research. In this workshop I will present the overview of the structure of the CSR and describe the peer review activities and the components of NIH peer review.

email: tdrgon@csr.nih.gov

NIH STATISTICAL METHODOLOGICAL GRANT APPLICATION AND REVIEW

Xihong Lin, Harvard School of Public Health*

I will discuss NIH statistical methodological grant writing and review process. Discussions will include both R01 grants for faculty and K99/R00 grants (Path to Independence (PI)) grants for postdoctoral fellows. Guidelines and experience will be shared.

email: xlin@hsph.harvard.edu

74. CAUSAL MEDIATION ANALYSIS: DEFINITIONS, IDENTIFICATION, INFERENCE AND CONTROVERSIES

ALTERNATIVE GRAPHICAL CAUSAL MODELS AND THE IDENTIFICATION OF DIRECT EFFECTS

Thomas Richardson, University of Washington*
James Robins, Harvard School of Public Health

We consider four classes of graphical causal models: the Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) of Robins (1986), the agnostic causal model of Spirtes et al. (1993), the Non-Parametric Structural Equation Model (NPSEM) of Pearl (2000), and the Minimal Counterfactual Model (MCM) which we introduce. The latter is referred to as "minimal" because it imposes the minimal counterfactual independence assumptions required to identify those causal contrasts representing the effect of an ideal intervention on any subset of the variables in the graph. The causal contrasts identified by an MCM are, in general, a strict subset of those identified by a NPSEM associated with the same graph. We analyze various measures of the "direct" causal effect, focussing on the pure direct effect (PDE), also called the "natural direct effect". We show the PDE is a parameter that may be identified in a DAG viewed as a NPSEM, but not as an MCM or FFRCISTG.

email: thomasr@u.washington.edu

WHY IS MEDIATION ANALYSIS NOT EASY?

Vanessa Didelez, University of Bristol, UK*

It is striking that in every day life as well as in applied research people readily and without much hesitation speak of direct and indirect effects in many different context, while a formal conceptualisation, criteria for identification and estimation of such effects outside of linear models seem surprisingly complicated. For example, if we want a notion that allows us to say that the total effect is the sum of direct and indirect effect then we have to use the natural/pure (in)direct effect, but at the same time it is debatable whether these are identifiable in any practical situation. I will discuss this issue against the background of a typical sociological question: the effect of mothers' education on the health of their children as possibly mediated by the duration of breastfeeding. I will further characterise types of questions and situations when mediation analysis is "easy".

email: vanessa.didelez@bristol.ac.uk

CAUSAL MEDIATION ANALYSIS FOR DICHOTOMOUS AND TIME-TO-EVENT OUTCOMES

Tyler VanderWeele, Harvard School of Public Health*

A key question in many studies is how to divide the total effect of an exposure into a component that acts directly on the outcome and a component that acts indirectly, i.e. through some intermediate. For example, one might be interested in the extent to which the effect of diet on blood pressure is mediated through sodium intake and the extent to which it operates through other pathways. In the context of such mediation analysis, even if the effect of the exposure on the outcome is unconfounded, estimates of direct and indirect effects will be biased if control is not made for confounders of the mediator-outcome relationship. Often data are not collected on such mediator-outcome confounding variables; the results in this paper allow researchers to assess the sensitivity of their estimates of direct and indirect effects to the biases from such confounding. Specifically, the paper provides formulas for the bias in estimates of direct and indirect effects due to confounding of the exposure-mediator relationship and of the mediator-outcome relationship. Under some simplifying assumptions, the formulas are particularly easy to use in sensitivity analysis. The bias formulas are illustrated by examples in the literature concerning direct and indirect effects in which mediator-outcome confounding may be present.

email: tvanderw@hsph.harvard.edu

SEMIPARAMETRIC THEORY FOR CAUSAL MEDIATION ANALYSIS: ROBUSTNESS, EFFICIENCY AND SENSITIVITY

Eric J. Tchetgen Tchetgen, Harvard University*
Ilya Shpitser, Harvard University

In recent years, scientists in the health sciences have become increasingly interested in mediation analysis. Specifically, upon establishing a non-null total effect of the exposure, investigators routinely wish to make inferences about the direct (indirect) pathway of the effect of the exposure not through (through) a mediator variable that occurs subsequently to the exposure and prior to the outcome. Although powerful semiparametric methodologies have been developed to analyze observational studies, that produce double robust and highly efficient estimates of the marginal total causal effect, similar methods for mediation analysis are currently lacking. Thus, we have developed a general semiparametric framework for obtaining inferences about so-called marginal natural direct and indirect causal effects, while appropriately accounting for a large number of pre-exposure confounding factors for the exposure and the mediator variables. Our analytic framework is particularly appealing, because it gives new insights on issues of efficiency and robustness in the context of mediation analysis. In particular, we propose new multiply robust locally efficient estimators of the marginal natural indirect and direct causal effects, and develop a novel double robust sensitivity analysis framework for the assumption of ignorability of the mediator variable.

email: etchetgen@gmail.com

75. ADVANCES IN BRAIN IMAGING AND SIGNAL BIOMARKERS FOR BEHAVIOR

HOW RESTFUL IS RESTING STATE fMRI? - A POPULATION FUNCTIONAL CHANGE-POINT ANALYSIS INVESTIGATION

John A.D. Aston, University of Warwick*
Claudia Kirch, Karlsruhe Institute of Technology

Resting state functional magnetic resonance imaging (fMRI) is a technique for examining the brain at rest. Here we examine whether the brain is active while at rest (in the sense of non-stationarities appearing in the time series) using change point detection in sequences of functional data. This is derived for situations where the spatial functional observations are temporally dependent and where the distributions of change points from multiple subjects is required. Of particular interest is the case where the change point is an epidemic change (a change occurs and then the observations return to baseline at a later time). The special case where the covariance can be decomposed as a tensor product is considered with particular attention to the power analysis for detection. This is of interest in fMRI where the estimation of a full covariance structure for the three-dimensional image is not computationally feasible. It is shown that in a large population of subjects (approx 200) around 50% exhibit detectable changes and the distribution of these changes is found.

email: j.a.d.aston@warwick.ac.uk

PREDICTING DISEASE STATUS USING A NOVEL SUPPORT VECTOR CLASSIFIER FOR LONGITUDINAL NEUROIMAGING DATA

F. DuBois Bowman, Emory University*
Shuo Chen, Emory University

An increasing number of neuroimaging studies are beginning to collect data longitudinally over different scanning sessions, for example before, during, and following treatment for a psychiatric disorder. Such studies may yield temporal changes in selected features that predictive of disease status or treatment response. Support vector machine (SVM) techniques are well-established tools that are applicable for classification and prediction of in high dimensional data settings. Current SVM methods, however, typically consider cross-sectional data collected during one time period or session (e.g. baseline). We propose a novel support vector classifier (SVC) for longitudinal high dimensional data that allows simultaneous estimation of the SVM separating hyperplane parameters and temporal trend parameters, which determine the optimal means to combine the longitudinal data for classification and prediction. We demonstrate the use and potential advantages of our proposed methodology using a simulation study and a data example from the Alzheimer's disease Neuroimaging Initiative. The results indicate that our proposed method leverages the additional longitudinal information to achieve higher accuracy than methods using only cross-sectional data and methods that combine longitudinal data by naively expanding the feature space.

email: dbowma3@emory.edu

DEVELOPING fMRI-BASED BIOMARKERS FOR PAIN

Martin A. Lindquist, Columbia University*

Biomarkers are a staple of medical tests, but progress in developing biomarkers for mental health-related phenomena has been slow. In this work we present a biomarker based on distributed patterns of fMRI activity that predicts physical pain based on normative data from other individuals. The same biomarker applied to a new study discriminated painful heat from non-painful warmth with high accuracy, but did not respond to social pain (i.e. rejection). The results indicate that physical and social pain are associated with different patterns of fMRI activity, even within regions commonly activated by both conditions. These results help establish a foundation for developing objective biomarkers of subjective phenomena related to mental health.

email: martin@stat.columbia.edu

NOVEL MEASURES OF DEPENDENCE IN TIME SERIES AS BIOMARKERS

Hernando Ombao, Brown University*
Mark Fiecas, Brown University
Cristina Gorrostieta, University of California at Irvine

We give an overview of approaches for analyzing dependence between brain regions. This project is motivated by a growing body of evidence suggesting that various neurological disorders, including Alzheimer's disease, depression, and Parkinson's disease may be associated with altered brain connectivity. In the first part of the talk, we shall discuss Granger-causality under the context of vector autoregressive models and then discuss the common spectral methods for characterizing dependence. In the second part of the talk, we shall discuss some open problems related to analyzing multivariate time series in an experimental setting. Here, we shall propose some general models that can capture transient features in each time series as well as between-trial and between-subject variability in the signals. This is joint work with PhD students Mark Fiecas and Cristina Gorrostieta and neuroscience collaborators at MGH.

email: ombao@stat.brown.edu

76. RECENT DEVELOPMENT IN IMPUTATION METHODS AND THEIR APPLICATIONS

A MULTIPLE IMPUTATION APPROACH TO MISREPORTING AND MISMEASUREMENT FROM MULTIPLE SOURCES

Yulei He, Harvard Medical School*
Mary Beth Landrum, Harvard Medical School
Alan Zaslavsky, Harvard Medical School

Measures of certain services variables (e.g., hospice use for patients with advanced stage) are important quality indicators in studies of patterns of cancer care, and can often be obtained from multiple sources. The Cancer Care Outcomes Research and Surveillance study collected these variables from several sources including a patient surrogate survey, medical records abstraction, and Medicare claims data. Variables subject to misreporting or mismeasurement can be both binary (has ever used hospice) and continuous (how many days before death if used). Yet important quality indicators are developed using all information (e.g., no hospice use or hospice use 3 or fewer days before death). Under a general assumption that none of the sources is the gold standard, we propose a multiple imputation approach to synthesizing data from all sources, incorporating the relationship among all variables and correcting the misreporting/mismeasurement. Valid analysis of the therapy variables can then be based on imputed/synthesized data.

email: he@hcp.med.harvard.edu

DOUBLY ROBUST NONPARAMETRIC MULTIPLE IMPUTATION FOR IGNORABLE MISSING DATA

Qi Long, Emory University*
Chiu-Hsieh Hsu, University of Arizona
Yisheng Li, University of Texas MD Anderson Cancer Center

Missing data are common in medical and social science studies and often pose a serious challenge in data analysis. We propose a new nonparametric multiple imputation (MI) approach for ignorable missing data that uses two working models to achieve dimension reduction and define the imputing sets for the missing observations. Compared with existing nonparametric imputation procedures, our approach can better handle covariates of high dimension, and is doubly robust in the sense that the resulting estimator remains consistent if either of the working models is correctly specified. Compared with existing semiparametric doubly robust methods, our nonparametric MI approach is more robust to the misspecification of both working models; it also avoids the use of inverse-weighting and hence is less sensitive to missing probabilities that are close to 1. We propose a sensitivity analysis for evaluating the validity of the working models, allowing investigators to choose the optimal weights so that the resulting estimator relies either completely or more heavily on the working model that is likely to be correctly specified and achieves improved efficiency. Our simulation studies show that the proposed method compares favorably with some existing methods in finite samples. The proposed method is further illustrated using data from a colorectal adenoma study.

email: qlong@emory.edu

WHY ARE THERE MULTIPLE HYPOTHESIS TESTING COMBINING RULES FOR MULTIPLY IMPUTED DATA SETS?

Xiao-Li Meng, Harvard University*
Xianchao Xie, Two Sigma Investments, LLC

Since the seminal work by Rubin (1987), multiple imputation has been extensively studied and widely applied in various areas. There is essentially only a single set of combining rules for point and interval estimation with multiply imputed data sets. However, the combining rules for hypothesis testing are of several types, depending whether we have access to complete-data point {and variance} estimators, test statistics, p-values, likelihood-ratio testing procedures, etc. And even within each type, currently it is largely unclear what is the optimal combining rule or even {how} to define the optimality. In this talk we explore such issues by first examining the "ideal" type, namely, performing Wald-type tests by combining complete-data point {and variance} estimates from individual imputed data sets. We propose a theoretically minor but practically useful modification to existing procedures as well as several variations, demonstrating that even for the current "ideal" combining rule, improvements are possible. We then investigate a number of new rules for combining complete-data p-values, based on a stochastic representation of the "ideal" Wald-type test. We conclude by discussing challenges for extending these improvements to multivariate cases and possible ways to overcome them.

email: meng@stat.harvard.edu

IMPUTING MODES FOR MISSING DATA BASED ON THE LAPLACE APPROXIMATION TO THE MARGINAL LIKELIHOOD

Myunghee Cho Paik, Columbia University*

Likelihood approaches involving missing data require the maximization of an integrated marginal likelihood, which can be a computationally challenging and labor-intensive task. Such a computational challenge often impedes the application of many useful models in practice. Laplace approximations have been widely used to evaluate marginal likelihood functions in Bayesian analysis and more recently in hierarchical likelihood, dubbed as the adjusted profile h-likelihood (APHL). Although the Laplace approximation is a mere computational tool used to maximize a marginal likelihood (as in the case of the EM algorithm), it possesses an intuitive appeal and simplicity, encouraging the dissemination of insightful models. In this talk, we survey a number of useful models where the APHL is valid and is shown to substantially alleviate computational burden, and inference on imputing value is possible.

email: mp9@columbia.edu

77. JOINT MODELING AND ITS APPLICATIONS

AN ESTIMATION METHOD OF MARGINAL TREATMENT EFFECTS ON CORRELATED LONGITUDINAL AND SURVIVAL OUTCOMES

Qing Pan, George Washington University
Grace Y. Yi, University of Waterloo*

This talk concerns treatment effects on correlated longitudinal and time to event processes. The marginal mean of the longitudinal outcome in the presence of event occurrence is often of interest from clinical and epidemiological perspectives. When the probability of the event is treatment dependent, differences between treatment-specific longitudinal outcome means are usually not constant over time. In this talk, we propose a measure to quantify treatment effects using time-varying differences in longitudinal outcome means, which accounts for the constantly changing population composition due to event occurrences. Generalized linear mixed models and proportional hazards models are employed to construct the proposed measure. The proposed method is applied to analyze the motivating data arising from the study of weight loss in the Diabetes Prevention Program where weights after diabetes occurrence are systematically different from diabetes-free weights.

email: yyi@uwaterloo.ca

A SEMIPARAMETRIC MARGINALIZED MODEL FOR LONGITUDINAL DATA WITH INFORMATIVE DROPOUT

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

We propose a marginalized joint-modeling approach for marginal inference on the association between longitudinal responses and covariates when longitudinal measurements are subject to informative dropouts. The proposed model is motivated by the idea of linking longitudinal responses and dropout times by latent variables while focusing on marginal inferences. We develop a simple inference procedure based on a series of estimating equations, and the resulting estimators are consistent and asymptotically normal with a sandwich-type covariance matrix ready to be estimated by the usual plug-in rule. The performance of our approach is evaluated through simulations and illustrated with a renal disease data application.

email: mengling.liu@nyu.edu

BAYESIAN SEMIPARAMETRIC NONLINEAR MIXED-EFFECTS JOINT MODELS FOR DATA WITH SKEWNESS, MISSING RESPONSES AND MEASUREMENT ERRORS IN COVARIATES

Yangxin Huang, University of South Florida*
Getachew A. Dagne, University of South Florida

It is a common practice to analyze complex longitudinal data using flexible semiparametric nonlinear mixed-effects (SNLME) models with the assumption of normality. Normality of model errors may be unrealistically obscuring important features of subject variations. To partially explain between- and within-subject variations, covariates are usually introduced in such models, but some covariates, however, may often be measured with substantial errors. Moreover, the responses may be missing and the missingness may be nonignorable. Inferential procedures can be complicated dramatically when data with skewness, missing observations and measurement errors are observed. In the literature, there has been considerable interest in accommodating either skewness, incompleteness or covariate measurement errors in such models, but there is relatively little work concerning all of the three features simultaneously. In this article, our objective is to address the simultaneous impact of skewness, missingness and covariate measurement errors by jointly modeling the response and covariate processes based on a flexible Bayesian SNLME model. The method is illustrated in a real AIDS data example to compare potential models with various scenarios and different distribution specifications.

email: yhuang@health.usf.edu

BAYESIAN HYBRID INFERENCE FOR LONGITUDINAL AND SURVIVAL JOINT MODELS

Gang Han, Moffitt Cancer Center & Research Institute*
Yangxin Huang, University of South Florida
Catherine Phelan, Moffitt Cancer Center & Research Institute

In longitudinal studies, often interest lies in the relation between longitudinally measured markers and a survival outcome, and thus it is desirable to model a longitudinal process and a survival process simultaneously. In statistical practice, usually either the frequentist or the Bayesian method is used in parametric inference. However, with a large number of parameters and limited data, the joint models might be non-identifiable. In order to solve this problem, some parameters with practical meaning and prior knowledge are better treated as Bayesian, while others such as nuisance parameters are better treated as frequentist. In this talk, we present a Bayesian hybrid approach to cope with two types of parameters in the longitudinal andsurvival joint models. Comparing with frequentist and Bayesian approaches, the proposed approach can be more accurate and efficient. We illustrate this approach in a HIV study or an ovarian cancer study.

email: gang.han@moffitt.org

JOINT SPATIAL MODELING OF RECURRENT INFECTION AND GROWTH IN FOREST ECOLOGY

Farouk S. Nathoo, University of Victoria*

We present new statistical methodology for longitudinal studies of disease ecology in forestry, where trees are subject to recurrent infection, and the hazard of infection depends on tree growth over time. Understanding the nature of this dependence has important implications for reforestation and breeding programs. Challenges arise for statistical analysis in this setting, with sampling schemes leading to panel data, exhibiting dynamic spatial variability, and incomplete covariate histories for hazard regression. In addition, data are collected at a large number of spatial locations which poses computational difficulties for spatiotemporal modeling. A joint model for infection and growth is developed; wherein, a mixed non-homogeneous Poisson process, governing recurring infection, is linked with a spatially dynamic nonlinear model representing the underlying height growth trajectories. These trajectories are based on the von Bertalanffy growth model and a spatially-varying parameterization is employed. Spatial variability in growth parameters is modeled through a multivariate spatial process derived through kernel convolution. Inference is conducted in a Bayesian framework with implementation based on Hamiltonian (Hybrid) Monte Carlo.

email: nathoo@math.uvic.ca

BAYESIAN JOINT MODEL OF MULTIVARIATE ORDINAL DATA WITH COMPETING RISKS SURVIVAL TIME

Satrajit Roychoudhury, Novartis Pharmaceuticals Corporation*

In longitudinal clinical trials missing data arises frequently in practice. Last observation carried-forward (LOCF) is used in most cases to handle missing values. But the LOCF method is likely to misrepresent the results of a trial. Alternative way to handle missing data is imputation via model based approaches. Model based approaches are gaining more and more interest in recent statistical literature. But existing joint models for longitudinal and survival data are not applicable for longitudinal ordinal outcomes with possible non-ignorable missing values caused by multiple reasons. We propose a semiparametric Bayesian joint model for longitudinal ordinal measurements and competing risks failure time data. In particular we develop a shared parameter model between the two endpoints and assumed a semiprametric Dirichlet process prior for the shared parameter. The robustness property of Dirichlet process (DP) allows automatic grouping of subjects and allow like-subjects to share information which results in improved parameter estimates. Proposed methodology is illustrated using data from clinical trial of intravenous recombinant tissue-plasminogen activator (rt-PA) in patients with acute stroke.

email: satrir@gmail.com

A JOINT LATENT CLASS MODEL OF SURVIVAL AND LONGITUDINAL DATA

Yue Liu, University of Virginia*
Lei Liu, University of Virginia
Jianhui Zhou, University of Virginia

There has been an increasing interest in the joint analysis of repeated measures and time to event data. In many studies, there could also exist heterogeneous subgroups. In this paper, we propose a new latent class model for the joint analysis of longitudinal and survival data. We use a latent class model to identify latent sub-populations. Within each latent class, we adopt a joint model of longitudinal and survival data. We apply our model to the Modification of Diet in Renal Disease (MDRD) Study. The result indicates two subtypes among the subjects when considering the longitudinal Glomerular Filtration Rate (GFR) measurements associated with time to death. Our model is desirable when the heterogeneity of subjects cannot be ignored and both the longitudinal and survival outcomes are of our interests.

email: yl7z@virginia.edu

78. BAYESIAN METHODS I

BAYESIAN KAPPA REGRESSION

Elande Baro, University of Maryland Baltimore County*
Zhen Chen, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Sung Duk Kim, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health
Bo Zhang, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Agreement between raters on disease diagnostics is typically assessed using the kappa coefficient. There has been considerable work using logistic regression to provide summary estimates of inter-rater agreement. Shoukri and Mian (1996) used covariates to predict the marginal probability of classification by each rater using a likelihood based approach. Klar et al (2000) used covariates to predict kappa using an estimating equations approach. We propose a method that identifies covariates predictive of kappa using a likelihood based approach under Bayesian framework. To illustrate this procedure, we study how patient age and physician gender affect inter-rater agreement in the diagnosis of the disease of endometriosis.

email: baroelande@hotmail.com

SPARSE DATA IN SAFETY DATA ANALYSES

Xiaowen Hu, Southern Methodist University*
Luyan Dai, Boehringer Ingelheim Pharmaceuticals
Tom Tang, Boehringer Ingelheim Pharmaceuticals

In safety data analyses of adverse events, sparseness often occurs. Many statistical methods are not valid or have problems in such situation. So how to handle sparseness, especially zero-event studies, is a big challenge. Quite a few methods have been proposed and there have been hot discussions on which one to use. However, no uniform agreement has been reached. This project explored several methods for calculating Risk Difference and Odds Ratio of Adverse Events, which are the two main metrics in safety data analyses, trying to identify their pros and cons under different situations. Some general guidance is obtained through extensive simulation studies. In the context of sparseness, Mantel-Haenszel Risk Difference overall has better performance among the three methods studied, while Peto's method is comparatively better among the four methods studied. Simulations of the Multivariate Bayesian Logistic Regression method through WinBUGS turned out to be very time-consuming; more efficient algorithm needs to be developed. Risk difference is additive measure while odds ratio is relative measure, which one to use depends on the clinical question at hand. Presenting risk difference alongside odds ratio is recommended.

email: christina125@gmail.com

MINKOWSKI-WEYL PRIORS FOR MODELS WITH PARAMETER CONSTRAINTS: AN ANALYSIS OF THE BIOCYCLE STUDY

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

We propose a general framework for performing full Bayesian analysis under linear inequality parameter constraints. The proposal is motivated by the BioCycle Study, a large cohort study of hormone levels of healthy women where certain well-established linear inequality constraints on the log-hormone levels should be accounted for in the statistical inferential procedure. Based on the Minkowski-Weyl decomposition of polyhedral regions, we propose a class of priors that are fully supported on the parameter space with linear inequality constraints, and we fit a Bayesian linear mixed model to the BioCycle data using such a prior. We observed a positive association between progesterone levels and F2-isoprostanes, a marker for oxidative stress, and a negative association between follicle stimulating hormone (FSH) and F2-isoprostanes. These findings are of particular interest to reproductive epidemiologists.

email: danahermr@mail.nih.gov

A PREDICTIVE BAYESIAN APPROACH TO THE DESIGN AND ANALYSIS OF BRIDGING STUDIES

A. Lawrence Gould, Merck Research Laboratories*
Jin Tian, Merck Research Laboratories
Li Xin Zhang, Merck Research Laboratories
William W. B. Wang, Merck Research Laboratories

Pharmaceutical product development culminates in confirmatory trials whose evidence for the product's efficacy and safety supports regulatory approval for marketing. Regulatory agencies in countries whose patients were not included in the confirmatory trials often require confirmation of efficacy and safety in their patient populations, which may be accomplished by carrying out "bridging studies" to establish consistency of the effects demonstrated by the original trials for local patients. We describe an approach for designing and analyzing 'bridging studies' that fully incorporates the information provided by the original trials. The approach determines probability contours of joint predictive intervals for treatment effect and response variability, or endpoints of treatment effect confidence intervals, that are functions of the findings from the original trials, the sample size for the "bridging study", and possible deviations from complete consistency with the original trials. A "bridging study" is judged consistent with the original trials if its findings fall within the probability contours; regulatory considerations determine the probability levels for the contours.

email: goulda@merck.com

BAYESIAN SEMIPARAMETRIC REGRESSION FOR EVALUATING PATHWAY EFFECTS ON ZERO INFLATED CLINICAL OUTCOMES

Lulu Cheng, Virginia Tech*
Inyoung Kim, Virginia Tech

In this study, we propose a semiparametric regression approach for identifying gene pathways related to zero inflated clinical outcomes. Our approach is developed by using a Bayesian hierarchial framework. We model nonparametrically pathway effect into a zero inflated Poisson hierarchical regression model with unknown link function. Nonparametric pathway effect was estimated via the kernel machine and the unknown link function was estimated by transforming a mixture of beta cumulative density function. Our approach provides flexible nonparametric settings to describe the complicated association between genes microarray expressions and the clinical outcomes. The Metropolis-within-Gibbs sampling algorithm and Bayes factor were adapted to make statistical inference. Our simulation results support that our semiparametric approach is more accurate and flexible than the ordinary zero inflated Poisson regression, this is especially so when the number of genes is large. The usefulness of our approaches is demonstrated through its applications to the Canine data set from Enerson et al. (2006). Our approaches can also be applied to other settings where a large number of highly correlated predictors are present.

email: lulu0220@vt.edu

BAYESIAN SAMPLING-BASED METHODS FOR INVERSE PREDICTION FROM LONGITUDINAL CD4 PROFILE DATA

Miranda L. Lynch, Harvard School of Public Health*
Victor DeGruttola, Harvard School of Public Health

We develop Bayesian methods for predicting time to reaching CD4 thresholds for antiretroviral treatment for HIV-infected treatment-naïve subjects whose current value of CD4 was measured in a cross-sectional household survey in Botswana. Information regarding rates of CD4 decline was obtained using information

from individuals with incident HIV infection followed over time in Botswana and South Africa. These data sources must be combined in order to estimate the additional amount of drugs that would be required to treat all such patients, should such treatment be started as soon as HIV infection is detected. To achieve this, we develop sampling based inverse prediction in a Bayesian framework that makes use of the model dependency structure in the auxiliary dependent dataset. The Bayesian methods allow proper characterization of variability in the predictions, as well as variability arising from using auxiliary data. We also discuss computational issues that arise in carrying out this sampling-based inverse prediction. These analyses permit assessment of costs associated with implementation of test-and-treat HIV prevention interventions.

email: mlynch@hsph.harvard.edu

ROBUST BAYESIAN INFERENCE FOR LONGITUDINAL MULTIVARIATE DATA WITH NORMAL/INDEPENDENT DISTRIBUTIONS

Sheng Luo, University of Texas at Houston*
Junsheng Ma, University of Texas at Houston
Karl D. Kieburtz, University of Rochester Medical Center
Barbara C. Tilley, University of Texas at Houston

Many randomized clinical trials have longitudinal multivariate outcomes measured in different scales, e.g. continuous, ordinal. We propose to use Item Response Theory (IRT) to evaluate the global treatment effects while accouting for all sources of correlation. Continuous outcomes are often assumed to be normally distributed, but this assumption is not robust against outlying observations. In this paper, we use three specific distributions, e.g. t-distribution, slash distribution, and contaminated normal distribution in normal/independent family to address the outlier issue. The models are compared and selected using deviance information criteria and Bayes factor.

email: junsheng.ma@uth.tmc.edu

79. CORRELATED / LONGITUDINAL DATA

A SEMIPARAMETRIC LATENT VARIABLE TRANSFORMATION APPROACH FOR MODELING MULTIPLE OUTCOMES OF MIXED TYPES

Anna Snavely, Harvard University*
Yi Li, Harvard University and University of Michigan

Often there are multiple correlated outcomes of varying types of interest in a given setting, rather than a single primary outcome. In the biomedical area, a failure time is frequently one of such outcomes. When we do have multiple outcomes, we would like to use all of the information provided in those outcomes in order to make some conclusion about a treatment or some other covariate. In this paper we propose a semiparametric latent variable normal transformation model that allows for the estimation of a treatment (or other covariate) effect in the presence of multiple outcomes, including failure times. Multiple outcomes are assumed to be governed by an unobserved (latent) variable, which in turn may depend on covariates such as treatment. As an extension of traditional latent variable approaches, our method allows the relationship between the outcomes and latent variables to be unspecified and allows for outcomes of mixed types which includes accounting for potentially censored outcomes. The method is applied to a study of head and neck cancer patients from Dana-Farber Cancer Institute in which multiple outcomes are available to characterize dysphagia.

email: asnavely@hsph.harvard.edu

ANALYSIS OF ASYNCHRONOUS LONGITUDINAL OBSERVATIONS

Hongyuan Cao, University of Chicago*
Donglin Zeng, University of North Carolina at Chapel Hill
Jason P. Fine, University of North Carolina at Chapel Hill

We consider nonparametric estimation in a generalized linear model for asynchronous longitudinal observations using estimating equations. The covariate is assumed to be generated from an underlying smooth function with univariate time index and the response can be very general-continuous, dichotomous or count data. Our results apply to the case where the covariate and response variable are not observed at the same time for each subject. We investigate both time-invariant and time-dependent coefficients estimation and their asymptotic properties. We illustrate our methods with an application to vital sign data and evaluate their finite-sample performance through simulation studies.

email: hycao@uchicago.edu

HIERARCHICAL MULTIPLE INFORMANT MODELS

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

Based on a non-standard approach of generalized estimating equation (GEE) methods, Pepe et al. (1999) and Horton et al. (1999) developed estimators for the association between univariate outcomes and multiple informant predictors. Their approach enables estimation of the marginal effect of each multiple informant predictor, and formal comparison among predictors in regard to the strength of their association with the outcome. We extend these multiple informant methods for hierarchical data structures to estimate and compare the strength of association among multiple correlated predictors while taking into account for the correlation within a cluster or a group. We applied the extended method to address two substantive questions regarding how features of the food environment near school affects child's body

mass index (BMI): 1) We investigate how the association between the number of fast food restaurants and child's BMI varies across several different buffer sizes from a school, and 2) compare the effect of two different features of the food environment (fast food restaurants and convenience stores). The newly developed methodology enhances the types of research questions that can be asked by investigators studying effects of environment on childhood obesity, although it can potentially be applied to other fields.

email: jongguri@umich.edu

MEASURES OF DISCRIMINATION FOR LATENT GROUP-BASED TRAJECTORY MODELS

Nilesh Shah, University of Pittsburgh*
Chung-Chou Chang, University of Pittsburgh

In clinical research, patient care decisions are often easier to make if patients are classified into a manageable number of groups based on homogeneous risk patterns. Investigators can use latent group-based trajectory modeling (Nagin, 2005) to estimate the posterior probabilities that an individual will be classified into a particular group of risk patterns. Although this method is increasingly used in clinical research, there is currently no measure that can be used to determine whether an individual's group assignment has a high level of discrimination. In this study, we propose a discrimination index and provide confidence intervals of the probability of the assigned group for each individual. We also propose a modified form of entropy to measure discrimination. The two proposed measures were applied to assess the group assignments of the longitudinal patterns of conduct disorders among pre-adolescent girls.

email: nhs3@pitt.edu

CHALLENGES IN ESTIMATION OF GENETIC EFFECTS FROM FAMILY-BASED CASE-CONTROL DATA

Roula Tsonaka, Leiden University Medical Center*
Jeanine J. Houwing-Duistermaat, Leiden University Medical Center

Estimation of genetic effects using case-control family data is often complicated by the outcome-dependent sampling and the within families correlation. An approach to deal with both of these features is the ascertainment corrected mixed-effects model. However, for small sample sizes and when the disease is rare convergence issues may arise, because the data do not provide sufficient information to estimate all model parameters. To overcome such numerical difficulties, it has proposed in the literature to combine data across different studies (Zheng etal, Biometrics, 2010). However, combining information from multiple sources is not always possible in practice. In this work, we propose an empirical Bayes method that introduces additional external information (e.g. disease prevalence) into the analysis of family data. Thereby, we can obtain reliable parameter estimates even for small samples and rare diseases. Finally, we illustrate our proposal using two studies on venous thrombosis: (i) the affected sibling pairs study GIFT, in which genotypes are available for cases and controls of sibships with at least two cases, and (ii) the case-control study MEGA, in which genotypes and family history is available for each case and control, while the genotypes of the relatives are missing.

email: s.tsonaka@lumc.nl

THE ANALYSIS OF CORRELATED NON-GAUSSIAN OUTCOMES FROM CLUSTERS OF SIZE TWO: NON-MULTILEVEL-BASED ALTERNATIVES?

Tom Loeys, Ghent University*
Geert Molenberghs, University of Leuven

In this presentation we discuss the analysis of clustered binary or count data, when the cluster size is two. For Gaussian outcomes, linear mixed models taking into account the correlation within clusters, are frequently used and well understood. Here we explore the potential of generalized linear mixed models (GLMMs) for the analysis of non-Gaussian outcomes that are possibly negatively correlated. Several approximation techniques (Gaussian quadrature, Laplace approximation or linearization) that are available in standard software packages for these GLMMs are investigated. Despite the different modelling options related to these different techniques, none of these have satisfactory performance in estimating fixed effects when the within-cluster correlation is negative and/or the number of clusters is relatively small. In contrast, a generalized estimating equations (GEE) approach for the analysis of non-Gaussian data turns out to have an overall excellent performance. When using GEE the robust score and Wald test are recommended for small and large samples, respectively.

email: tom.loeys@ugent.be

CONDITIONAL INFERENCE FUNCTIONS FOR MIXED-EFFECTS MODELS WITH UNSPECIFIED RANDOM-EFFECTS DISTRIBUTION

Peng Wang, Bowling Green State University*
Guai-feng Tsai, U.S. Food and Drug Administration
Annie Qu, University of Illinois at Urbana-Champaign

In longitudinal studies, mixed-effects models are important for addressing subject-specific effects. However, most existing approaches assume a normal distribution for the random effects, and this could affect the bias and efficiency of the fixed-effects estimator. Even in cases where the estimation of the fixed effects is robust with a misspecified distribution of the random effects, the estimation of the random effects could be invalid. We propose a new approach to estimate fixed and random effects using conditional quadratic inference functions. The new approach does not require the specification of likelihood functions or a normality assumption for random effects. It can also accommodate serial

correlation between observations within the same cluster, in addition to mixed-effects modeling. Other advantages include not requiring the estimation of the unknown variance components associated with the random effects, or the nuisance parameters associated with the working correlations. Real data examples and simulations are used to compare the new approach with the penalized quasi-likelihood approach, and SAS GLIMMIX and nonlinear mixed effects model (NLMIXED) procedures.

email: wangp@bgsu.edu

80. IMAGING

A BAYESIAN HIERARCHICAL FRAMEWORK FOR MODELING BRAIN CONNECTIVITY OF NEUROIMAGING DATA

Shuo Chen, Emory University*
F. DuBois Bowman, Emory University
Lijun Zhang, Emory University

The challenges for modeling brain connectivity based on neuroimaging data lie in (i) ultrahigh dimensionality: one typical neuroimaging data consists tens of billions of voxel pair connectivities per subject, (ii) complex hierarchy such as voxel, region, and population levels, (iii) interaction of connectivities of neural unit pairs within the same neural circuitry. To respond those challenges, we propose a novel Bayesian hierarchical framework that unifies voxel level, region level, and population level brain connectivity analysis. The first level of the proposed method summarizes voxel pair level brain connectivities between a region pair by two subgroups: connected and non-connected voxel pairs by using mixture model. Based on the "small-worldness" property , we use the proportion of connected voxel pairs between two regions as the region pair level connectivity strength metric and then investigate the clinical covariates at population level. Additionally, we provide a solution to account for the the covariance of region pair level connectivities. The posteriors enable inferences on whole brain voxel and region level connectivities as well as clinical covariate effects such as age, gender and treatment groups that may impact connectivities. We apply our method to a example data set and simulation study.

email: chenshuochen@gmail.com

SIMPLE MODIFICATIONS OF A t-TEST FOR IMPROVED POWER WITH FDR CONTROL IN fMRI

Shuzhen Li, Medtronic, Inc.
Lynn E. Eberly, University of Minnesota*
Brian S. Caffo, Johns Hopkins University

For signal detection in fMRI data, statistical tests are often constructed across participants at each voxel. The test may compare, for example, brain activation during a task to brain activation during no-task, or the task vs. no-task difference in activation between patients and matched controls. fMRI studies are expensive, typically resulting in small sample sizes, thus the voxel-specific error variance estimates are inefficient. This is a promising context for methods which borrow information across voxels. We propose a new approach for the borrowing of information, and compare it to several established methods: the usual t-statistic with no shrinkage, a moderated t-statistic proposed by Smyth (2004), and a James-Stein shrinkage t-statistic by Cui et al. (2004). The new approach is based on Efron (2007) where the null distribution of the test statistic is estimated using quantile transformed normal distributions and data truncation. We instead propose building a likelihood based on the t distribution and data truncation. False discovery rate (FDR) thresholding is then applied to each of the resulting p-value maps (one for each statistical approach) to conduct statistical inference. The criteria of true FDR and sensitivity are compared to demonstrate the performance of our methods in large simulations. An example face recognition task is presented.

email: eberl003@umn.edu

ADAPTIVE THRESHOLDING FOR fMRI DATA

Joke Durnez, Ghent University, Belgium*
Beatrijs Moerkerke, Ghent University, Belgium

When analyzing functional MRI-data, several thresholding procedures are available to account for the huge number of volume units or features that are tested simultaneously. The main focus of these methods is to prevent an inflation of false positives. However, this comes with a serious decrease in power and leads to a problematic imbalance between type I and type II errors. In this research, we present a method to estimate the number of activate features. More precisely we consider peaks of activation as a topological feature. Knowledge of the number of active features has two important implications: (1) Widely used methods to control the false discovery rate (FDR) can be made adaptive and more powerful, and (2) using thresholding procedures, the type I and type II error rate can be estimated enabling a direct trade-off between sensitivity and specificity. The method is evaluated and illustrated using simulations and a real data example.

email: joke.durnez@ugent.be

APPLICATION OF CLUSTER ANALYSIS IN DEMENTIA RESEARCH

Jay Mandrekar, Mayo Clinic*

Research in academic medical centers offer opportunities to collaborate on clinical projects that involve novel applications of statistical methods. Frontotemporal dementia is a diverse group of uncommon disorders of the brain that primarily affect the frontal and temporal lobes. These are the areas of the brain that are usually associated with personality, behavior and language. This talk will focus on applying cluster analysis techniques to data from frontotemporal dementia. The goal is to assign similar observations to smaller groups called clusters, such that observations within clusters are similar to each other than those in other clusters.

email: mandrekar.jay@mayo.edu

THREE-DIMENSIONAL RECOGNITION OF STEM CELLS USING AN ENTROPY BASED NONPARAMETRIC HYPOTHESIS TESTING APPROACH

Ran Liu, University of Connecticut*
Dipak K. Dey, University of Connecticut

Stem cells have many potential applications in new medical treatments, such as diabetes, heart disease and Parkinson's disease. The automated stem cell recognition and classification are of great importance in developing such medical treatments. In this paper, we develop a nonparametric hypothesis testing procedure to test the equality of entropy from different samples, following Gangopadhyay, Disario, and Dey (1997) for the recognition and classification of embryonic stem cells (ES cells) and fibroblast stem cells (FB cells). The three-dimensional stem cell holographic image data are obtained through digital holographic microscopy, which provides both magnitude and phase information of the stem cells. We apply the procedure to test the equality of entropy of the holographic images of the ES cells and FB cells. An effective algorithm is also developed on the discriminant analysis of ES cells and FB cells.

email: ranliu84@gmail.com

A BAYESIAN APPROACH TO DETERMINING FUNCTIONAL CONNECTIVITY IN THE HUMAN BRAIN WITH INCORPORATION OF STRUCTURAL CONNECTIVITY

Wenqiong Xue, Emory University*
F. DuBois Bowman, Emory University

Recent innovations in neuroimaging technology have provided opportunities for researchers to investigate the anatomy as well as the function of the human brain. Investigating functional and structural relationships in the brain stands to improve our understanding of neural networks and the pathophysiology of these networks underlying psychiatric and neurologic disorders. We present a unified Bayesian framework for analyzing functional connectivity of the human brain utilizing the knowledge of associated structural connections. Our functional connectivity measure rests upon assessments of functional coherence, which is based on simultaneously elevated regional brain activity levels. Our structural connectivity information is drawn from probabilistic diffusion tensor tractography (DTT), which is a technique applied to diffusion tensor imaging data that quantifies probabilities of structural connectivity. We formulate a prior distribution for functional coherence between brain regions that depends upon the probability of structural connectivity between the regions, with this dependence adhering to structure-function links revealed by fMRI and DTT data. Posterior estimation is performed using Markov Chain Monte Carlo (MCMC) techniques implemented via Gibbs sampler and the Metropolis algorithm.

email: wxue@emory.edu

81. LONGITUDINAL AND TIME SERIES DATA ANALYSIS

STATE-SPACE TIME SERIES CLUSTERING USING DISCREPANCIES BASED ON THE KULLBACK-LEIBLER INFORMATION AND THE MAHALANOBIS DISTANCE

Eric D. Foster, University of Iowa*
Joseph E. Cavanaugh, University of Iowa

Time series applications frequently arise in biomedicine, genetics, and bioinformatics that require the clustering of multiple series into homogeneous groups. Both nonparametric and parameteric techniques have been formulated; the latter are often based on discrepancy measures developed within a suitable modeling framework. For the purpose of clustering state-space processes, Bengtsson and Cavanaugh (2008) proposed the use of a discrepancy based on a Kullback-Leibler information measure. This measure is derived using the joint distribution of the collection of smoothed values for the states, computed via the Kalman filter smoother. In this work, we formulate a Mahalanobis distance version of the joint Kullback-Leibler based discrepancy, so as to focus more on the trajectory of any given time series than the corresponding process. We also propose and investigate analogous discrepancies based on composite measures derived from the marginal distributions of the smoothers. For comparison purposes, we contrast these measures to counterparts derived using the observed series as opposed to the smoothed series. In addition, we consider a Euclidean distance that does not require any modeling. Our simulation results indicate that the measures based on the smoothed series outperform those based on the observed series.

email: eric-foster@uiowa.edu

DEVELOPMENTAL TRAJECTORIES OF MARIJUANA USE FROM ADOLESCENCE TO ADULTHOOD: PERSONALITY AND SOCIAL ROLE OUTCOMES

Judith S. Brook, NYU School of Medicine
Jung Yeon Lee, NYU School of Medicine*
Elaine N. Brown, NYU School of Medicine
Stephen J. Finch, State University of New York, Stony Brook
David W. Brook, New York University School of Medicine

Longitudinal trajectories of marijuana use from adolescence into adulthood were examined for adverse life-course outcomes among African-Americans and Puerto Ricans. Data for marijuana use were analyzed at four points in time and on participants' personality attributes, work functioning, and partner relations in adulthood using growth mixture modeling. Each of the three marijuana-use trajectory groups (maturing-out, late-onset, and chronic marijuana-users) had greater adverse life-course outcomes than a nonuse or low-use trajectory group. The chronic marijuana-use trajectory group was highly associated with criminal behavior and partners' marijuana use in adulthood. Treatment programs for marijuana use should also directly address common adverse life-course outcomes users may already be experiencing.

email: jungyeon.lee@nyumc.org

MODELING THE EVOLUTION OF NEUROPHYSIOLOGICAL SIGNALS

Mark Joseph A. Fiecas, Brown University*
Hernando Ombao, Brown University

We develop a new procedure for nonidentical nonstationary bivariate time series data which has two sources of nonstationarity: 1) within each replicate (or within a trial of a neuroscience experiment) and 2) across the replications (across trials). Thus, the spectral properties of the time series are evolving over time within a replicate and also over the replications. We propose a novel statistical model and corresponding two-stage estimation method for estimating the spectral properties of the time series data that takes into account these two sources of nonstationarity. In the first stage we account for nonstationarity over time within a replicate using local periodogram matrices. In the second stage, we account for nonstationarity over the replications using wavelet regression by pooling the wavelet coefficients obtained from "neighboring" replicates to obtain smoother estimates. We illustrate our approach on a simulated data set and then apply the method to a local field potential (LFP) data to study how the cross-coherence between the nucleus accumbens and the hippocampus evolves over the course of a learning association experiment.

email: mfiecas@stat.brown.edu

MARKOV REGRESSION MODELS FOR COUNT TIME SERIES WITH EXCESS ZEROS: A PARTIAL LIKELIHOOD APPROACH

Ming Yang, University of Iowa*
Gideon Zamba, University of Iowa
Joseph Cavanaugh, University of Iowa

Count data with excess zeros are common in many biomedical and public health applications. The zero-inflated Poisson (ZIP) regression model has been widely used in practice to analyze such data. In this paper, we extend the classical ZIP regression framework to model count time series with excess zeros. A Markov regression model is presented and developed, and the partial likelihood is employed for statistical inference. Partial likelihood inference has been successfully applied in modeling time series where the conditional distribution of the response lies within the exponential family. Extending this approach to ZIP time series poses methodological and theoretical challenges, since the ZIP distribution is a mixture and therefore lies outside the exponential family. Under the partial likelihood framework, we devise the EM algorithm to compute the maximum partial likelihood estimator (MPLE). We establish the asymptotic theory of the MPLE under mild regularity conditions and investigate its finite sample behavior in a simulation study. The performances of different model selection criteria are compared in the presence of model misspecification. Finally, we present an epidemiological application to illustrate the proposed methodology.

email: ming-yang@uiowa.edu

SEMIPARAMETRIC APPROACH TO A NON-LINEAR RANDOM EFFECTS QUANTILE REGRESSION MODEL

Mi-Ok Kim, Cincinnati Children's Hospital Medical Center*
Rhonda Vandyke, Cincinnati Children's Hospital Medical Center

We consider a non-linear random effects quantile regression analysis of clustered data and propose a semiparametric approach using empirical likelihood. The random regression coefficients are assumed independent with a common mean, following parametrically specified distributions. We formulate the estimation of the random coefficients as an estimating equations problem and use empirical likelihood to incorporate the parametric likelihood of the random coefficients. A likelihood-like statistical criterion function is yield. We use this -like statistical criterion function as the likelihood and Markov Chain Monte Carlo (MCMC) samplers in the Bayesian framework. We propose the resulting quasi-posterior mean as an estimator. We illustrate the methodology with a real data example of the Cystic Fibrosis (CF) Foundation Patient Registry data. Age related lung function decline is of research interest and random effects quantile regression methodology enables quantile specific rates of decline conditioning on age, accommodating variation across CF centers.

email: miok.kim@cchmc.org

BUILDING A NEW CONTROL CHART FOR BIOSURVEILLANCE

Yiying Fan, Cleveland State University*

Development of new methods of statistical process control (SPC) is extremely important for modern bio-surveillance applications. Typical challenges in SPC are that the data are correlated and multivariate. This study introduces a new joint control chart and provides a general approximation to the joint distribution of average and maximum of a continuous time process. The new control chart is motivated from combining the merits of Shewhart and CUSUM charts for process monitoring. We use continuous Gaussian processes with given covariance functions and discrete autoregressive moving average (ARMA) processes to evaluate the new control chart for both in-control and out-of-control performances in comparison to the standard methods. It is shown through simulation that the new method is efficient and compares well to the standard control methods.

email: richard_fan90@yahoo.com

ROBUST ESTIMATION OF MIXED EFFECTS MODEL FOR FINITE NORMAL MIXTURES

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

In this work, we develop robust estimation approach for the analysis of clustered data with multimodal conditional distributions. It is proposed to model such data in a hierarchical model with conditional distributions viewed as finite mixtures of normal components. With a large number of observations in the lowest level clusters, a two-stage estimation approach is used. In the first stage, the normal mixture parameters in each lowest level cluster are estimated using robust minimum divergence estimators. This robust alternative to the maximum likelihood estimation is used to provide stable results even for data with conditional distributions such that components may not quite meet normality assumptions. Then suitably transformed lowest level cluster-specific means and standard deviations are modeled in a linear mixed effects (LME) model in the second stage. Robust M-estimation is used for the second stage LME model. The proposed modeling approach is illustrated through the analysis of mice tendon fibril diameters data.

email: tingtingzhan@gmail.com

82. SURVIVAL ANALYSIS AND RISK PREDICTION

PARTLY CONDITIONAL ESTIMATION OF THE EFFECT OF A TIME-DEPENDENT FACTOR IN THE PRESENCE OF DEPENDENT CENSORING

Qi Gong, University of Michigan*
Douglas E. Schaubel, University of Michigan

We propose semiparametric methods for estimating the effect of a time-dependent covariate on treatment-free survival. The data structure of interest consists of a longitudinal sequence of measurements and a potentially censored survival time. The factor of interest is time-dependent. Treatment-free survival is of interest and is dependently censored by the receipt of treatment. Patients may be removed from consideration for treatment, temporarily or permanently. The proposed methods involve landmark analysis and partly conditional hazard regression. Dependent censoring is overcome by Inverse Probability of Censoring Weighting (IPCW). The predicted quantities are marginal in the sense that time-varying covariates are taken as fixed at each landmark, with the mortality hazard function implicitly averaging across future covariate trajectories. The proposed methods circumvent the need for explicit modeling of the longitudinal covariate process. The proposed estimators are shown to be consistent and asymptotically normal, with consistent covariance estimators provided. Simulation studies reveal that the proposed estimation procedures are appropriate for practical use. We apply the proposed methods to pre-transplant mortality among End-stage Liver Disease (ESLD) patients.

email: gongqi@umich.edu

REGRESSION ANALYSIS OF CLUSTERED INTERVAL-CENSORED FAILURE TIME DATA WITH THE ADDITIVE HAZARDS MODEL

Junlong Li, University of Missouri*
Chunjie Wang, Mathematics School and Institute of Jilin University
Jianguo Sun, University of Missouri

This paper discusses regression analysis of clustered failure time data, which means that the failure times of interest are clustered into small groups instead of being independent and often occur in many fields such as medical studies. For the problem, a number of methods have been proposed, but most of them apply only to clustered right-censored data. In reality, the failure time data are often interval-censored. That is, the failure times of interest are known only to lie in certain intervals. We proposes an estimating equation-based approach for regression analysis of clustered interval-censored failure time data generated from the additive hazards model. A major advantage of the proposed method is that it does not involve the estimation of any baseline hazard function. Both asymptotic and finite sample properties of the proposed estimates of regression parameters are established and the method is illustrated by the data arising from a lymphatic filariasis study.

email: junlong.li@mail.missouri.edu

LANDMARK RISK PREDICTION OF RESIDUAL LIFE FOR BREAST CANCER SURVIVAL

Layla Parast, Harvard University*
Tianxi Cai, Harvard University

The importance of developing personalized risk prediction estimates has become increasingly evident in recent years. In general, patient populations may be heterogenous and represent a mixture of different unknown subtypes of disease. When the source of this heterogeneity and resulting subtypes of disease are unknown, accurate prediction of survival may be difficult. However, in certain disease settings the onset time of an observable short term event may be highly associated with these unknown subtypes of disease and thus may be useful in predicting long term survival. One approach to incorporate short term event information along with baseline markers for the prediction of long term survival is through a landmark Cox model, which assumes a proportional hazards model for the residual life at a given landmark point. In this paper, we use this modeling framework to develop procedures to assess how a patient's long term survival trajectory may change over time given good short term outcome indications along with prognosis based on baseline markers. We first propose time-varying accuracy measures to quantify the predictive performance of landmark prediction rules for residual life and provide resampling-based procedures to make inference about such accuracy measures. Simulation studies show that the proposed procedures perform well in finite samples. Throughout, we illustrate our proposed procedures using a breast cancer dataset with information on time to metastasis and time to death. In addition to baseline clinical markers available for each patient, a chromosome instability genetic score, denoted by CIN25, is also available for each patient and has been shown to be predictive of survival for various types of cancer. We provide procedures to evaluate the incremental value of CIN25 for the prediction of residual life and examine how the residual life profile changes over time. This allows us to identify an informative landmark point such that accurate risk predictions of the residual life could be made for patients who survive past the landmark point without metastasis.

email: lparast@hsph.harvard.edu

ESTIMATING RESTRICTED MEAN JOB TENURES FOR COMPENSATORY DAMAGES IN PROMOTION DISCRIMINATION CASES: APPLICATION TO ALEXANDER VS. MILWAUKEE

Qing Pan, George Washington University*
Joseph Gastwirth, George Washington University

Besides coefficient estimates, researchers are often interested in predictions from survival processes. The issue is further complicated when multiple processes are influential on the outcome. In estimating the compensatory damages in equal employment cases, both promotion and retirement processes affect the length of time a subject serves on different ranks; the hypothetical job tenures for different salary levels without discrimination is of interest. We treat the promotion and retirement as the event and termination processes respectively. Estimators of restricted mean life time for the status before promotion, after promotion and after retirement are proposed respectively. The asymptotic distribution of the restricted mean life time estimator is derived and examined through simulation studies. The robustness of the predictions in the presence of frailty terms are evaluated. The method is applied to reverse discrimination case Alexander vs. Milwaukee. The corresponding expected lengths of time serving as a lieutenant, serving as a captain and being retired are estimated as references for compensation.

email: qpan@gwu.edu

STATISTICAL METHODS OF TIME-CONDITIONAL SURVIVAL

Victoria Gamerman, University of Pennsylvania*
Phyllis A. Gimotty, University of Pennsylvania

For cancer patients interested in their current prognosis, traditional survival statistics reported from time of diagnosis are no longer relevant. Time-conditional survival estimates use conditional probability as an alternative measure of future survival by accounting for time elapsed from diagnosis. Over the last two decades, clinical investigators have presented point estimates and corresponding 95% confidence limits for time-conditional survival probabilities. We develop the asymptotic distribution for log time-conditional survival and use weighted least squares to fit regression models to the log time-conditional survival probabilities as a function of time survived after diagnosis. Hypothesis tests comparing various models to the saturated model are developed to address the clinically relevant questions. Esophageal cancer survival is used to illustrate the proposed methodology.

email: vica@mail.med.upenn.edu

100 YEARS ON: A NEW LOOK AT SURVIVORSHIP ON THE TITANIC

Stephen D. Walter, McMaster University*
Hedy Jiang, McMaster University
Corinne A. Riddell, McGill University

The unsinkable Titanic sank almost 100 years ago off the coast of Newfoundland, on April 15, 1912, en route to New York City. Fifteen hundred lives were lost and about seven hundred persons survived. Several previous studies have shown that passengers' chances of surviving the sinking were related to their age, sex and the class in which they travelled. Here we investigated the effects of other social and economic factors, including the number of travelling companions, occupations, birthplaces and residences of passengers and crew. We created an extensive dataset, based on the information available from Encyclopedia Titanica, a source that claims to have among the most accurate passenger and crew lists ever compiled. Standardized death rates were calculated for various groups of interest, using the indirect method adjusting for age and class. The impact of these factors was also explored using multiple logistic regression models. Finally, we examined the patterns of death among crew members, a group that has received relatively little attention in the literature.

email: walter@mcmaster.ca

ADJUSTED SURVIVAL ANALYSIS WITH INVERSE PROBABILITY WEIGHTS IN COMMUNITY-BASED PRIMARY CARE PRACTICES

Zugui Zhang, Christiana Care Health System*
Edward Ewen, Christiana Care Health System
Paul Kolm, Christiana Care Health System

Without random assignment of treatment, selection bias is often introduced in observational studies when the controls are not representative of the study base. Results from survival analysis in observational studies could be misleading due to confounding. Patients with better prognostic values were more likely to be assigned to the new treatment at baseline; a higher survival rate could be found in the treated group than in the untreated group. The purpose of this study was to apply adjusted survival analysis with inverse probability weights to examine the approach to antithrombotic therapy and patterns of warfarin use, the quality of anticoagulation, and their relationship with subsequent stroke and bleeding events in a series of community-based primary care practices. Data of 1141 patients were obtained from an electronic medical record encompassing office and hospital care, in which fifty-five stroke cases were identified and bleeding is also another outcome. Age and gender were the two primary risk factors. Interruptions in warfarin use were the major exposure condition. A CHADS2 score, the best validated clinical prediction rule for determining the risk of stroke, was used to adjust for analysis as well. Adjusted survival models with stabilized inverse probability weights were applied to assess the independent effect of warfarin interruptions on stroke and bleeding.

email: zzhang@christianacare.org

83. STATISTICAL METHODS AND APPLICATIONS IN RARE VARIANT SEQUENCING STUDIES

JOINT MOMENT TEST FOR RARE VARIANTS

Daniel J. Schaid, Mayo Clinic*

Rare variants are likely to have a prominent role in the etiology of complex traits, yet there are significant statistical challenges to model their effects on traits. If all rare variants in a gene have the same direction of effect on a trait, then the popular "burden" test, based on a weighted sum of the variants within a gene, provides reasonable power. In contrast, if variants are a mixture of both risk and protective effects, then the alternative C(alpha) test (for case-control data), a test for binomial over-dispersion, can have greater power. These two strategies can be viewed as first-moment and second moment-moment tests, respectively, which are nearly orthogonal. The power of each strategy depends on the mixture of risk and protective effects. We propose a novel strategy that combines both moment tests into a joint simultaneous test of first and second moments, based on generalized linear models. This provides a flexible way to incorporate covariates, and to extend to other types of traits (e.g., quantitative). By using a simultaneous test of both moments, we gain robustness to compensate for our lack of knowledge about the balance of protective, neutral, and risk variants. Theory and simulations will be presented to illustrate the power advantages of different strategies to test for associations of rare variants with traits.

email: schaid@mayo.edu

A NOVEL PERMUTATION STRATEGY TO CORRECT FOR CONFOUNDERS IN CASE-CONTROL STUDIES OF RARE VARIATION

Michael P. Epstein, Emory University*
Richard Duncan, Emory University
Yunxuan Jiang, Emory University
Karen N. Conneely, Emory University
Andrew S. Allen, Duke University
Glen A. Satten, Centers for Disease Control and Prevention

Many case-control tests of rare variation [Neale et al. PLoS Genete1001322; Ionita-Laza et al. PLoS Genet e1001289; Li et al. AJHG 87: 728 among others] are implemented in statistical frameworks that prohibit straightforward correction for confounders such as population stratification. To correct for this confounding, we propose establishing the significance of a rare-variant test using a novel permutation procedure that preserves the confounding present within the sample. Using Fisher's noncentral hypergeometric distribution, we sample disease outcomes for subjects in a permuted dataset in a manner such that the probability a subject is selected as a case is dependent on his/her odds of disease conditional on confounder variables. Using both simulated sequence data as well as real data from the Dallas Heart Study, we demonstrate that our permutation strategy corrects for confounding due to population stratification that, if ignored, would inflate the size of a rare-variant test. The permutation approach is applicable to any rare-variant association test used in a case-control study and is implemented in a modified version of the R package "BiasedUrn" for public use.

email: mpepste@emory.edu

INVESTIGATING THE IMPACT OF THE RARE SPECTRUM OF VARIATION ON DRUG REPOSITIONING AND DRUG RESPONSE

Matthew R. Nelson, GlaxoSmithKline*

Over the past five years, our capacity to measure common variants in the human genome have led to many important discoveries affecting drug response. There are several examples of common variants having a large effect on adverse drug reactions and drug efficacy, some that have been translated into common medical practice. However, little is known about

the contribution of rare genetic variation to drug response, in particular for rare adverse reactions known to be under genetic influence. Advances in sequencing technologies are opening up opportunities to investigate the role of rare genetic variation for all forms of complex human traits, which will impact drug discovery and development. There are several large-scale sequencing experiments being conducted to assess the impact of rare variants on drug repositioning opportunities and drug response, yielding some early insights into this area. We will share important lessons learned through these early sequencing studies and discuss implications for future study design and analysis.

email: matthew.r.nelson@gsk.com

KERNEL MACHINE BASED TESTING OF RARE VARIANT BY ENVIRONMENT INTERACTIONS

Michael C. Wu, University of North Carolina at Chapel Hill*

Rare variants are believed to strongly influence complex traits. Accordingly, a plethora of computational methods have been developed for data processing and testing main rare variant effects. However, despite increasing interest, little methodological work has been done on identifying rare variant by environment interactions, yet such methods are keenly needed by many studies examining gene-environment interactions. Therefore, we propose a statistical framework for region based analysis of rare variant by environment interactions by extending and tailoring the flexible kernel machine regression framework to test the cumulative environmental and rare variant interaction effect across multiple rare variants within a region and multiple environmental exposures. Specifically, we choose to model the environmental effects, the rare variant effects, and their interactions using an additive kernel machine regression model. Significance is evaluated via a score test exploiting connections between kernel machines and mixed models. We develop a weighting strategy for incorporating prior biological and bioinformatics annotation information and allele frequency. Simulations and real data analysis establish that our kernel machine based interaction test is a powerful tool for identifying novel gene by environment interactions in sequencing studies.

email: mwu@bios.unc.edu

84. CAUSAL INFERENCE METHODS FOR HIV RESEARCH

PRACTICAL APPLICATIONS OF PRINCIPAL STRATIFICATION IN HIV RESEARCH

Bryan E. Shepherd, Vanderbilt University*

Principal stratification approaches have been proposed for dealing with post-randomization selection. However, these methods have been criticized for, among other things, focusing on the wrong questions. I will discuss some applications in HIV research and critically examine to what extent these methods focus on relevant scientific questions.

email: bryan.shepherd@vanderbilt.edu

ESTIMATION OF JOINT EFFECTS OF MULTIPLE TIME-VARYING EXPOSURES IN INFECTIOUS DISEASE RESEARCH

Stephen R. Cole, University of North Carolina at Chapel-Hill*
Chanelle J. Howe, Brown University

Frequently in infectious diseases research, the primary interest is in the joint effects of two or more time-varying treatments or exposures. Use of joint marginal structural models to consistently estimate such effects remain largely absent from the applied literature a decade after their introduction by Hernán, Brumback and Robins (JASA 2001) perhaps due to a paucity of compelling examples. We provide two worked examples both studying HIV acquisition. In example 1, using data on 1,525 African-American adults in the AIDS Link to Intravenous Experience cohort study we estimated the joint effects of alcohol consumption and intravenous drug use. In example 2, using data on 3,725 men in the Multicenter AIDS Cohort Study we estimated the joint effects of alcohol consumption and the number of partners with whom unprotected receptive anal intercourse was practiced. More widespread use of joint marginal structural models is needed to estimate the effects of multiple time-varying exposures in infectious disease research.

email: cole@unc.edu

MEDIATION ANALYSIS FOR OBSERVATIONAL EVENT TIME DATA

Jing Zhang, Brown University
Joseph W. Hogan, Brown University*
Catherine Gichunge, Moi University
Edwin Sang, Moi University
Abraham Siika, Moi University

Many statistical methods for mediation analysis have been developed for randomized trials. In many large scale HIV programs, particularly in the developing world, evaluating the process by which new interventions effect changes in outcome is important.

■ = Student Award Winner | ★ = Presenter

We develop methods for mediation analysis for observational event time data. The methods are used to analyze data from the USAID-funded AMPATH program. Our analysis investigates whether food aid reduces mortality by increasing adherence to scheduled clinic visits for HIV positive individuals who initiate antiviral therapy.

email: jhogan@stat.brown.edu

85. MODERN STATISTICAL MACHINE LEARNING FOR COMPLEX AND HIGH DIMENSIONAL DATA

HIGH-DIMENSIONAL PHARMACOEPIDEMIOLOGY

David Madigan, Columbia University*

Regulators approve drugs as safe end effective on the basis of clinical trials. Such trials necessarily provide limited insights and much attention now focuses on harnessing large-scale high-dimensional healthcare databases to better understand approved drugs. This talk will describe recent developments in this area.

email: madigan@yahoo.com

HDLSS DISCRIMINATION WITH ADAPTIVE DATA PILING

Myung Hee Lee, Colorado State University
Jeongyoun Ahn, University of Georgia*
Yongho Jeon, Yonsei University

We propose new discrimination methods for classification of High Dimension, Low Sample Size (HDLSS) data which regularize the degree of data piling. The within-class scatter of the HDLSS data, when projected onto a low dimensional discriminant subspace, can be selected to be arbitrarily small. Utilizing this fact, we develop two different ways of tuning the amount of within-class scatter, or equivalently, the degree of data piling. In the first approach we consider a linear path connecting the maximal data piling and the least data piling directions. We also formulates a problem of finding the optimal classifier under a constraint on data piling. The data piling regularization methods are extended to the multi-category problems. Simulated and real data examples show competitive performances of the proposed classification methods.

email: jyahn@uga.edu

LIKELIHOOD ADAPTIVE MODIFIED PENALTY AND ITS PROPERTIES

Tengfei Li, Fudan University
Yang Feng, Columbia University*
Zhiliang Ying, Columbia University

For variable selection, balancing sparsity and stability is a very important task. In this work, we propose the Likelihood Adaptive Modified Penalty (LAMP) where the penalty function is adaptively changed with the type of the likelihood function. Two types of asymptotic stability are defined and it is shown that LAMP can achieve both types of stability while achieving oracle properties. In addition, LAMP could be seen as a special functional of a conjugate prior. An efficient coordinate-descent algorithm is proposed and a balancing method is introduced. Simulation results and real data analysis show LAMP has competitive performance comparing with several well-known penalties.

email: yangfeng@stat.columbia.edu

REGULARIZED MULTIPLE-INDEX MODEL FOR GROUP VARIABLE SELECTION

Sijian Wang, University of Wisconsin-Madison*

In many biological applications, there is a natural groupings of predictors. For example, assayed genes or proteins can be grouped by biological roles or biological pathways. Traditional variable selection methods tend to make selection based on the strength of individual variables rather than the strengths of groups of variables, and may have inadequate variable selection or prediction performances. In this talk, we propose a regularized multiple-index model to integrate group structures to the model. It can not only identify important groups, but also select important predictors within selected groups. Furthermore, the proposed method has three good properties: 1) It allows a flexible modeling on the association between response and predictors which may yield better prediction performance; 2) It considers the interactions among variables within the same group; 3) When the groups have overlaps, i.e., one predictor can belong to several group, it can distinguish the effects of a predictor in all of groups it belongs to. We analyze several TCGA cancer datasets to demonstrate the proposed method.

email: swang@biostat.wisc.edu

86. STATISTICAL CHALLENGES IN REPRODUCTIVE AND ENVIRONMENTAL EPIDEMIOLOGY

CONCEPTUAL & METHODOLOGIC CHALLENGES UNDERLYING THE ASSESSMENT OF ENVIRONMENTAL REPRODUCTIVE AND DEVELOPMENTAL TOXICANTS: AN OVERVIEW

Germaine M. Louis, National Institute of Child Health and Development, National Institutes of Health*

Human reproduction and development encompasses a series of highly timed and interrelated processes including both partners of the couple and with several endpoints difficult to measure at the population level (e.g., ovulation, conception, embryonic quality). With an evolving body of evidence suggesting that endocrine disrupting chemicals may be potential reproductive and/or developmental toxicants, the need for sensitive methods and analytic techniques to assess chemical mixtures in relation to reproductive and developmental sensitive endpoints is paramount. This talk will present an overview of the methodologic nuances introduced by human reproductive and developmental processes to stimulate methods development suitable for population based research.

email: louisg@mail.nih.gov

MODELING TIME-TO-PREGNANCY IN TERMS OF VARIABILITY OF MENSTRUAL LENGTH

Amita Manatunga, Emory University*
Huichao Chen, Harvard University
Limin Peng, Emory University
Michele Marcus, Emory University

We adopt a proportional odds (PO) model to investigate the influences of potential fertility predictors on time-to-pregnancy. Compared to the traditional discrete proportional risk model, the PO model may provide more straightforward interpretations, while having the same capacity of handling right censoring and left truncation of TTP data. To address the particular interest in the effect of subject-specific menstrual cycle variability, which is not directly available from the data, on fecundability, we propose a sensible quantity to summarize within-woman variability in menstrual cycle length (MCL) and model its association with covariates via a log-linear model. A natural two-stage procedure is developed to estimate the coefficients for demographic variables of interest and that for cycle variability under the PO model. We establish the consistency and asymptotic normality of the resulting estimators. Simulations demonstrate good finite-sample performance of our proposals. Finally we present a generalization of the PO model to accommodate cycle-specific effects on fecundability. We illustrate our methods by an application to the study of Mount Sinai Study of Women Office Workers (MSSWOW).

email: amanatu@emory.edu

ANALYSIS OF IN-VITRO FERTILIZATION DATA WITH MULTIPLE OUTCOMES USING DISCRETE TIME TO EVENT ANALYSIS

Arnab Maity, North Carolina State University*
Paige Williams, Harvard School of Public Health
Louise Ryan, Commonwealth Scientific and Industrial Research Organisation
Stacey Missmer, Harvard School of Public Health
Brent Coull, Harvard School of Public Health
Russ Hauser, Harvard School of Public Health

In vitro fertilization (IVF) is an increasingly common method of assisted reproductive technology. IVF studies provide an ideal opportunity to identify and assess clinical and demographic factors along with environmental exposures that may impact IVF success rates. The main challenge in analysis of data resulting from IVF studies is the presence of multiple hierarchically-ordered outcomes per individual, resulting from both multiple opportunities for pregnancy loss within a single IVF cycle in addition to multiple IVF cycles. To date, most evaluations of IVF studies do not make use of full data due to its complex structure. In this paper, we develop statistical methodology for analysis of IVF data with multiple cycles and possibly multiple failure types observed for each individual. We develop a general methodology based on a generalized linear modeling formulation that allows implementation of various types of models including shared frailty models, failure specific frailty models, and transitional models. We apply our methodology to an analysis of the data from IVF study conducted by the Brigham and Women's Hospital, Massachusetts. We also summarize the performance of our proposed methods based on a simulation study.

email: amaity@ncsu.edu

BAYESIAN BORROWING OF INFORMATION ACROSS HIGH-DIMENSIONAL EXPOSURES AND OUTCOMES

Amy H. Herring, University of North Carolina at Chapel Hill*
David B. Dunson, Duke University
Andrew F. Olshan, University of North Carolina at Chapel Hill

Many birth defects are too rare to be studied in isolation. To facilitate analysis, birth defects are often "lumped" into larger groups based on the organ system affected or developmental origins. However, investigators are concerned that "lumping" may obscure important relationships that would be apparent if power permitted splitting defects into finer groups. We discuss methods for borrowing of information and shrinkage across high-dimensional environmental, biomedical, pharmacological, and sociodemographic risk factors, many of which are too rare to be studied in isolation. We show results based on studying environmental exposures and birth defects in the National Birth Defects Prevention Study, one of the largest studies of the causes of birth defects ever conducted.

email: aherring@bios.unc.edu

87. COMBINING POPULATION DATA FROM MULTIPLE SOURCES

COMBINING INFORMATION FROM MULTIPLE COMPLEX SURVEYS

Qi Dong, University of Michigan*
Trivellore Raghunathan, University of Michigan
Michael Elliott, University of Michigan

This article describes the use of multiple imputation to combine information from multiple surveys of the same underlying population. The basic proposal is to simulate synthetic populations from which the respondents of each survey have been selected. In this process, different sampling designs of the multiple surveys will be taken into account. Once we have the synthetic populations, we could treat them as simple random samples with no complex sampling design features and borrow information across surveys to adjust for nonsampling errors or fill in the variables that are lacking in one or more surveys. Then, we can analyze each synthetic population with standard complete-data software for simple random samples and obtain valid inference by combining the point and variance estimates first across synthetic populations within each survey using the existing combining rules for synthetic data and then across multiple surveys using the methods developed in this article. A model-based method to produce the synthetic populations is discussed and evaluated. It is shown that, by borrowing the information across multiple surveys using the methods in this article, we obtain more accurate and precise estimates for the statistics of interest.

email: qidong@umich.edu

ESTIMATING EFFECTIVENESS OF HEALTH CARE COMBING INFORMATION FROM DIFFERENT SURVEYS

Trivellore Raghunathan, University of Michigan
Irina Bondarenko, University of Michigan*

Effectiveness of medical care can be expressed as improvement in national health in return for the dollars spent for medical care. Obviously, the three questions arise: (1) How to measure national health? (2) How to measure the cost of medical care attributed to various diseases? and (3) How does then one relate answers to questions (1) and (2) to estimate quantities useful for health policy decisions. To answer these questions we need to link medical expenditure to diagnostic, treatment and preventative strategies, as well as trends in cost allocation and disease prevalence. The proposed framework allows us to combine information from multiple surveys, where one includes a variable of interest observed only on the subset of its plausible values, whereas the other survey has complete data. We explore strategies to attribute cost to the specific diseases based on multiple data sets. The method is applied to the Medicare Current Beneficiary Survey (MCBS) for years 1999-2004 to estimate cost attributable to various health conditions. The national Health and Nutrition Examination Survey (NHANES) plays the role of the "gold standard" for disease prevalence.

email: ibond@umich.edu

LONGITUDINAL ANALYSIS OF LINKED DATA: A CASE STUDY

Guangyu Zhang, National Center for Health Statistics*
Jennifer Parker, National Center for Health Statistics
Nathaniel Schenker, National Center for Health Statistics

Record linkage is a very valuable and efficient tool for connecting information from different data sources. The data linkage program at the National Center for Health Statistics (NCHS) combines data from population health surveys with data from administrative records, including records from the Centers for Medicare and Medicaid Services (CMS), expanding the usefulness of both sources of information and adding longitudinal information to cross-sectional survey data. In this research, we study mammography screening using the National Health Interview survey (NHIS, year 2004-2005) linked to the CMS (Medicare) administrative data (year 1999-2007). Several complicated issues are addressed statistically, such as unlinked records and nonresponse, correlations within subjects (due to repeated measurements) and between subjects (due to the complex sample design), and unbalanced data in the form of varying numbers of Medicare data years available for each survey respondent. Simulation studies are conducted to evaluate our approach.

email: nschenker@cdc.gov

COMBINING DATA FROM PROBABILITY AND NON-PROBABILITY SURVEYS

Michael R. Elliott, University of Michigan*
Alexa Resler, University of Michigan
Carol Flannagan, University of Michigan
Jonathan Rupp, University of Michigan

Analysts may want to combine data from probability and non-probability surveys for multiple reasons, including the fact that the non-probability sample may contain the detailed outcomes of interest, the non-probability sample may be substantially larger than the probability sample, and fellow analysts may demand to use non-probability samples in lieu of probability samples in many settings. Indeed, non-probability samples are likely increasing as Web surveys become increasingly entrenched in market research and other settings. Survey methodologists arguably should propose methods that can improve the quality of analyses obtained from these datasets, at least under clearly specified assumptions. Here we develop a method we term "pseudo-weighting" to reduce or remove bias in under model assumption that can be tested, at least in part. We present a set of simulations, along with an application combining data from NASS-CDS, a probability sample of automobile crashes, and CIREN, a non-probability sample of emergency room admission from automobile crashes.

email: mrelliot@umich.edu

88. SPATIAL UNCERTAINTY IN PUBLIC HEALTH PROBLEMS

SPATIAL UNCERTAINTY IN HEALTH DATA: DOES IT MATTER AND WHY SHOULD I WORRY ABOUT IT?

Geoffrey Jacquez, Biomedware*

Geocoding Positional Errors Can Impact Disease Rates; Disease Clustering; Odds Ratios; and Estimates of Individual Exposures, Resulting in Flawed Findings and Poor Public Health Decisions. Yet Geocoding Positional Error is Often Ignored, Primarily Because of a Lack of Theory, Methods and Tools for Propagating Positional Errors. Important Knowledge Gaps Include a Detailed Understanding of Empirical Geocoding Error Distributions and their Spatial Autocorrelation Structure; Impacts of Positional Error on the Statistical Power of Spatial Analysis Methods; Models for Predicting Positional Error at Specific Locations; and the Propagation of Positional Errors Through Health Analyses. A Research Agenda is Proposed to Address Five Key Needs. A Lack Of: 1) Standardized, Geocoding Resources for Use in Health Research; 2) Validation Datasets that will Allow the Evaluation of Alternative Geocoding Procedures; 3) Spatially Explicit Geocoding Positional Error Models; 4) Resources for Assessing the Sensitivity of Spatial Analysis Results to Positional Error; 5) Demonstration Studies of the Sensitivity of Health Policy Decisions to Geocoding Positional Accuracy.

email: jacquez@biomedware.com

RELATING PUBLIC HEALTH TO ENVIRONMENTAL FACTORS: QUANTIFYING UNCERTAINTY WHEN EXPOSURE IS PREDICTED

Linda J. Young, University of Florida*
Kenneth K. Lopiano, University of Florida
Carol A. Gotway, U.S. Centers for Disease Control and Prevention

Publicly available data from disparate sources are increasingly combined for subsequent statistical analyses. Because the data are frequently measured or associated with different geographic or spatial units, combining them for analysis usually requires prediction of one or more of the variables of interest. For example, health outcomes and related covariates may be measured at residences (points) or reported only at the zip code or county level while environmental exposure is measured at monitors (points). In either case, the support for health and environmental measurements differ. To assess the association between the two, environmental exposure is predicted for the points or areal units for which health outcomes are observed. When exposure is predicted using a smoothing method, such as kriging, Berkson error arises in the estimation of the regression parameters relating the predicted environmental exposure to health outcomes. If the kriging parameters must be estimated, classical measurement error is also introduced. In this talk, methods for accounting for both Berkson and classical measurement error are discussed. Whether and when it is important to account for Berkson and classical measurement error are considered. Data arising from Florida's Environmental Public Health Tracking Program are used to illustrate.

email: LJYoung@ufl.edu

SPATIAL UNCERTAINTY AND SPATIAL MEASURES OF PERFORMANCE

Lance Waller, Emory University*

We consider aspects of spatial uncertainty relating to statistical measures of model performance (e.g., power, sensitivity, and specificity). In particular we review general global measures of statistical performance relating to goodness of fit, probabilities of detection, and false alarm rates, then place these in a spatial perspectives. The spatially heterogeneous nature of study populations often leads to wide variation in performance associated to local sample sizes, neighborhood definitions, and the nature of summary statistics. We illustrate this concept with examples relating to the detection of spatial clusters of disease, spatial data fusion, and summaries of parametric model fitting. While these features are largely features of most spatial data, the examples illustrate how a spatial perspective for model evaluation plays an important part in spatial analysis.

email: lwaller@emory.edu

VISUALIZING STATISTICS AND UNCERTAINTY PATTERNS WITH MICROMAPS

Daniel B. Carr, George Mason University*
Linda W. Pickle, StatNet Consulting LLC

Micromaps are graphics that use an organized set of small maps to help in representing statistics. Micromap designs vary depending on the task involved. The task may include exploring or communicating patterns related to associations among variables, spatial indices and sometimes temporal indices. Measures or indicators of uncertainly are readily incorporated in the three major classes of micromap designs. Linked micromaps examples from http://statecancerprofiles.cancer.gov/micromaps/ routinely include confidence intervals that provide one way to represent estimate uncertainly. Dynamically conditioned choropleth maps (CCmaps) can easily use confidence interval widths as one of two conditioning variables. Comparative micromaps, which are designed to address change blindness when comparing map series, can indicate regions where statistics are suppressed due to data quality or confidentially considerations. Such a series can be compared with another map series showing population size or other variable potentially related to the uncertainty. Current micromaps provide ways to represent uncertainty in a spatial context and more ways may emerge as micromap designs evolve.

email: dcarr@gmu.edu

89. NEW STATISTICAL TOOLS FOR HIGH DIMENSIONAL PROBLEMS

DISCOVERING GRAPHICAL GRANGER CAUSALITY IN SPARSE HIGH-DIMENSIONAL NETWORKS WITH INHERENT GROUPING STRUCTURE

George Michailidis, University of Michigan*
Sumanta Basu, University of Michigan
Ali Shojaie, University of Washington

The problem of estimating high-dimensional network structures observed over time arises naturally in the analyses of many physical, biological and socio-economic systems. Examples include stock price fluctuations in financial markets and gene regulatory networks representing effects of regulators (transcription factors) on regulated genes in Genetics. We aim to learn the structure of the network over time employing the framework of Granger causal models under the assumptions of sparsity of its edges and inherent grouping structure among its nodes. We introduce a truncated penalty variant of Group Lasso to discover the Granger causal interactions among the nodes of the network. Asymptotic results on the consistency of the new estimation procedure are developed. The performance of the proposed methodology is assessed through an extensive set of simulation studies and comparisons with existing techniques.

e-mail: gmichail@umich.edu

QUANTILE REGRESSION IN ULTRA-HIGH DIMENSION

Lan Wang, University of Minnesota*
Yichao Wu, North Carolina State University
Runze Li, The Pennsylvania State University

We advocate a more general interpretation of sparsity which assumes that only a small number of covariates influence the conditional distribution of the response variable given all candidate covariates; however, the sets of relevant covariates may differ when we consider different segments of the conditional distribution. In this framework, we investigate the methodology and theory of nonconvex penalized quantile regression in ultra-high dimension. The proposed approach has two distinctive features: (1) it enables us to explore the entire conditional distribution of the response variable given the ultra-high dimensional covariates and provides a more realistic picture of the sparsity pattern; (2) it requires substantially weaker conditions compared with alternative methods in the literature; thus, it greatly alleviates the difficulty of model checking in the ultra-high dimension. In theoretic development, it is challenging to deal with both the nonsmooth loss function and the nonconvex penalty function in ultra-high dimensional parameter space. We introduce a novel sufficient optimality condition, which enables us to establish the oracle property for sparse quantile regression in the ultra-high dimension under relaxed conditions. The proposed method greatly enhances existing tools for ultra-high dimensional data analysis.

e-mail: wangx346@umn.edu

STATISTICAL TOOLS FOR IDENTIFYING AND PREDICTING MULTIPLE PATHWAYS

Joseph S. Verducci, The Ohio State University*
Samuel Handelman, The Ohio State University
Steven Bamattre, The Ohio State University

Biological adaptation often takes several forms, and specific traits may be associated with hundreds of genetic polymorphisms. Examples include cancers becoming chemo-resistant and viruses becoming more transmittable. Two new statistical tools are being developed to help investigate such phenomena: Multi-tau-path to discover distinct subpopulations with different forms of association; and PhyloPTE (Phylogenetic Path To Event) to find likely mutational paths to trait development.

e-mail: verducci.1@osu.edu

SELECTING THE NUMBER OF PRINCIPAL COMPONENTS IN FUNCTIONAL DATA

Yehua Li, University of Georgia*
Naisyin Wang, University of Michigan
Raymond J. Carroll, Texas A&M University

We consider functional data measured at discrete time points and contaminated with measurement error. Based on asymptotic theory, we propose two information-based criteria. The first criterion is built on marginal modeling approach and it is consistent. However, if one is willing to impose some structural assumptions, then a modified Akaike information criterion (AIC) we propose tends to obtain much improved numerical performance in choosing the correct number of principal components than the marginal and other existing criteria. This modified AIC is based on the expected Kullback-Leibler information under a conditional Gaussian-distribution modeling consideration. Our framework covers both sparse and dense functional data. Finite sample performance of the proposed information criteria is illustrated through simulation studies, where we show that our methods vastly outperform previously proposed criteria. An empirical example on colon carcinogenesis data is also provided to illustrate the results.

e-mail: yehuali@gmail.com

ROBUST ESTIMATION OF LARGE GAUSSIAN GRAPHICAL MODEL

Peng Tang, Georgia Institute of Technology
Huijing Jiang, IBM T.J. Watson Research Center
Xiwnei Deng, Virginia Tech*

The covariance matrix and its inverse have drawn increasingly attentions recently in many research areas such as bioinformatics. A desirable covariance matrix estimate is expected to be robust to outliers. The robustness of the estimate is a critical issue in

high-dimensional data analysis. In this work, we proposed a robust inverse covariance matrix estimation using the integrated squared error as a loss function. By imposing L1 regularization, the proposed estimate gains the sparse representation for the graphical model as well as the estimation robustness. The performance of the proposed method is illustrated through some simulation study and a real application in gene expression network.

e-mail: xdeng@vt.edu

90. BAYESIAN METHODS II

A BAYESIAN APPROACH FOR RANK AGGREGATION

Ke Deng, Harvard University*
Xuxin Liu, Harvard University
Jiong Du, Peking University
Jun S. Li, Harvard University

Rank aggregation, i.e., combining several ranked lists obtained from different sources to get a consensus, is an important problem in many disciplines. Most methods in literature assume that the different ranked lists are homogeneous in terms of reliability, and treat them equally. However, it is very common that the given ranked lists have diverse qualities, i.e., some of them are more reliable than the others. In this talk, we introduce a novel Bayesian method for rank aggregation problem that allows us to model the quality variations of the data.

e-mail: dengke3@gmail.com

BAYESIAN INFERENCE FOR CASE-CONTROL STUDIES WITH MULTIPLE NON-GOLD STANDARD EXPOSURE ASSESSMENTS: WITH AN APPLICATION IN OCCUPATIONAL HEALTH

Jing Zhang, University of Minnesota*

In many occupational case-control studies, work-related exposure assessments are often error-prone measurements of the true underlying exposure. In the absence of a gold standard, two or more imperfect assessments are often used to assess exposure. In this condition, misspecification of non-differential or differential misclassification, and independent or dependent misclassification conditioning on the latent exposure status, will often lead to biased estimation of the exposure-disease association. Although methods have been proposed to study diagnostic accuracy in the absence of a gold standard, these methods are infrequently used in case-control studies to simultaneously consider dependent and differential misclassification. In this paper, we proposed a Bayesian method to estimate the measurement-error corrected exposure-disease association, accounting for both differential and dependent misclassification. The performance of the proposed method is investigated using simulations, as well as an application to a case-control study assessing the association between asbestos exposure and mesothelioma.

e-mail: jingzhang2773691@gmail.com

A NONPARAMETRIC BAYESIAN MODEL FOR LOCAL CLUSTERING

Juhee Lee, University of Texas MD Anderson Cancer Center*
Peter Mueller, University of Texas at Austin
Yuan Ji, University of Texas MD Anderson Cancer Center

We propose a nonparametric Bayesian local clustering (NoB-LoC) approach for heterogeneous data. Using genomics data as an example, the NoB-LoC clusters genes into gene sets and simultaneously creates multiple partitions of samples, one for each gene set. Inference is guided by a joint probability model on all random elements. Posterior probabilities are reported for the random clustering of genes and for the nested random partition of samples with respect to each gene set. Biologically, the model formalizes the notion that biological samples cluster differently with respect to different genetic processes, and that each process is related to only a small subset of genes. These local features are importantly different from global clustering approaches such as hierarchical clustering, which create one partition of samples that applies for all genes in the data set. Furthermore, the NoB-LoC includes a special cluster of genes that do not give rise to any meaningful partition of samples. These genes could be irrelevant to the disease conditions under investigation. Similarly, for a given gene set, the NoB-LoC includes a subset of samples that do not co-cluster with other samples. The samples in this special cluster could, for example, be those whose disease subtype is not characterized by the particular gene set.

e-mail: jlee14@mdanderson.org

A BAYESIAN CHARACTERIZATION FOR A WEIGHTED SUM OF ENVIRONMENTAL CHEMICALS

Stephanie M. Pearson, Virginia Commonwealth University*
Roy T. Sabo, Virginia Commonwealth University

Exposures to endocrine disrupting chemicals (EDC) can consist of many chemicals, from different sources with different effects on hormonal function. The use of whole-mixture approaches, where a set of chemical concentrations is transformed into one representative value, are often prohibited or negatively influenced by the lack of toxic equivalence factors (TEQs) or relative potencies. Methods that ignore these values can mask the relationships between individual chemicals and some biomarker. A Bayesian approach to estimating a weighted sum of the chemical concentrations is presented that consists of a whole-mixture methodology that maintains information on individual chemicals. Bayes methods are used to specify prior distributions for the chemical-specific weights, while standard methods are used to characterize prior information for regression and variance parameters. MCMC methods are used to establish posterior inference on the relationship between the weighted mixture sum and an endocrine response. This weighted-sum approach achieves parametric parsimony with respect to the number of regression estimates, yet captures the contributions of individual chemicals to the overall mixture relationship via the estimated weights. Simulation studies compare this methodology with standard whole-mixture approaches.

e-mail: pearsonsm@vcu.edu

ESTIMATING REPRODUCTIVE INHIBITION POTENCY IN AQUATIC TOXICITY TESTING WHEN EXCESS ZEROS OBSERVED

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

Effectively and accurately assessing the toxicity of chemicals and their impact to the environment continues to be an important concern in ecotoxicology. Single experiments conducted by a particular laboratory often serve as the basis of a toxicity assessment. When the testing organisms are exposed to higher toxicity concentrations, the mortality rates usually increases and the resulting number of total young would include a certain proportion of zeroes. In this paper, a Bayesian analysis for potency estimation based on a single experiment was formulated, which then served as the basis for incorporating the historical control experimental information in application studies. A Bayesian hierarchical model was developed to estimate the reproductive inhibition concentrations (RIp) of a toxin and handle the excess zeroes. The methods were illustrated using a data set produced by the Ceriodaphnia dubia reproduction test in which the number of young produced over three broods is recorded. In addition, simulation studies were included to compare the Bayesian methods with previously proposed potency estimators when different number of animals were used in the experiments. The Bayesian methods gave more precise RIp estimates with smaller variation and nominal coverage probability.

e-mail: zhangj8@muohio.edu

BAYESIAN ARMITAGE-DOLL MULTISTAGE CARCINOGENESIS MODEL IN ESTIMATING CANCER MORTALITY

Zhiheng Xu, Emory University*
Vicki Hertzberg, Emory University

Although the etiology of cancer remains under investigation, evidence has suggested that multiple events occur during carcinogenesis, the process of the transformation of normal cells into cancer cells. The Armitage-Doll multistage model has been successfully employed in many carcinogenesis studies due to its simplicity and success in predicting cancer mortality rate. The model provides estimates of different numbers of stages for various types of cancer by assuming the death rate due to cancer proportional to age at death raised to a power that is one less than the number of stages between normal health and death. Biologically this variation by type reflects the different pathophysiological mechanisms leading to different cancer outcomes. In this paper, we first employed an alternative Bayesian approach in the Armitage-Doll multistage model to fit different types of cancer mortality data. Different likelihoods and prior settings were discussed and sensitivity analysis and model assessment showed that the Bayesian Armitage-Doll model fits the cancer mortality data w*ell.*

e-mail: zxu4@emory.edu

91. DIAGNOSTIC AND SCREENING TESTS

DISCRETE SURVIVAL ANALYSIS WITH MISCLASSIFIED EVENTS

Abidemi Adeniji, University of Pittsburgh*

Kaplan-Meier (KM) product limit estimator (Kaplan EL, Meier P. JASA 53:457-81, 1958) is the most commonly used method to estimate the survival function, specifically, in the presence of censoring. Numerous studies have employed this method assuming that the outcome is known with certainty. However, unless the event of interest is terminal (such as death), clinical diagnosis of an event is often subjected to misclassification, where the outcome is given with some uncertainty. In the presence of diagnostic errors, the true survival distribution of the time to first event is unknown. We estimate the true survival distribution by incorporating negative predictive values and positive predictive values of the prediction process. This will allow us to quantify the bias in the KM survival estimates due to the presence of misclassified events in the observed data. We present an unbiased estimator of the true survival rates and its variance. Asymptotic properties of the proposed estimators are provided analytically and these properties are examined through simulations. We demonstrate our methods using a sample of the dataset from the VIRAHEP-C study.

e-mail: abk38@pitt.edu

A NEW APPROACH TO ADJUST FOR VERIFICATION BIAS IN ASSESSMENT OF BINARY DIAGNOSTIC TESTS

Qingxia Chen, Vanderbilt University*

Verification bias can occur in diagnostic assessment when verification of disease status depends on the result of the screening test and/or patient characteristics. A patient not verified by the gold standard can be viewed as missing the value of the true disease status. In this article, the accuracy measurements of screening tests, including positive/negative predictive values, sensitivity, and specificity, are estimated based on a pseudo likelihood. Our estimates use propensity score and predictive mean score to condense multi-dimensional auxiliary covariates so that missing at random assumption is likely to be valid. The proposed estimates are shown to possess the doubly robust property, i.e. they are consistent when the model for the disease status or the model for the verification status, both conditioning on the auxiliary covariates, is correct. A number of simulations are used to compare the numerical performance between our estimates and other existing methods in the literature. Sensitivity analysis is conducted to evaluate the missing at random assumption. A real data is used to illustrate the methods.

e-mail: cindy.chen@vanderbilt.edu

DIAGNOSTIC TESTS BASED ON MULTIPLE CUTPOINTS FOR NOT PROPER ROC CURVES

Peter R. Dawson, University of Pennsylvania*
Phyllis A. Gimotty, University of Pennsylvania

In ROC curve analysis, the clinical utility of a continuous biomarker is often assessed by creating a binary diagnostic test defined by a cutpoint. Selection of an optimal cutpoint is commonly driven by maximizing the sensitivity and specificity of the diagnostic test and can be determined through a variety of measures including Youden's Index. When the relationship between the continuous biomarker and the probability of disease is not monotone, the ROC curve will lie on both sides of the random chance line. Selection of a single cutpoint is not optimal for such biomarkers. Instead we propose to create an optimal diagnostic test using two cutpoints which allows for better classification of patients into diseased and healthy classes. We examine two methods for selecting the pair of cutpoints. The first compares all possible pairs of cutpoints using Youden's Index and the second more efficiently selects two cutpoints based on the distance between the ROC curve and the random chance line. We report on simulation studies assessing the performance of both proposed algorithms as well as comparing our new diagnostic tests to the traditional single cutpoint approach. Finally, we applied the methods to non-monotone biomarkers identified from a cancer gene expression dataset.

e-mail: dawsonp@mail.med.upenn.edu

ESTIMATION OF THE VOLUME UNDER THE ROC SURFACE WITH THREE ORDINAL DIAGNOSTIC CATEGORIES USING KERNEL SMOOTHING

Le Kang, University at Buffalo*
Lili Tian, University at Buffalo

With three ordinal diagnostic categories, one of the most commonly used measures for the overall diagnostic accuracy is the volume under the receiver operating characteristic (ROC) surface (VUS), which is the extension of the area under the ROC curve (AUC) for binary diagnostic outcomes. In this talk, we discuss and compare different procedures for estimation of this summary index of diagnostic accuracy, namely, VUS. These estimation procedures are based on (i) the Mann-Whitney U statistic; (ii) the empirical plug-in estimator; (iii) normality assumption; (iv) kernel smoothing; and (v) Box-Cox type transformations to normality. As different estimation procedures would provide different estimated VUS, it is therefore of critical importance to examine their properties. We compare these estimation procedures in terms of bias and root mean square error in an extensive simulation study to provide recommendation of which approach is to be preferred in different scenarios.

e-mail: lekang@live.com

SOFT ROC CURVES

Yixin Fang, New York University
Narayanaswamy Balakrishnan, McMaster University
Xin Huang, Fred Hutchinson Cancer Research Center*

Receiver operating characteristic (ROC) curves are a popular tool for evaluating continuous diagnostic tests. However, the traditional definition of ROC curves incorporates implicitly the idea of "hard" thresholding, which cannot encompass the situation when some intermediate classes are introduced between test result positive and negative, and also results in the empirical curves being step functions. For this reason, we introduce here the definition of soft ROC curves, which incorporates the idea of "soft" thresholding. The softness of a soft ROC curve is controlled by a regularization parameter that can be selected suitably by a cross-validation procedure. A byproduct of the soft ROC curves is that the corresponding empirical curves are smooth. The methods developed here are then examined through some simulation studies as well as a real illustrative example.

e-mail: xhuang@fhcrc.org

EVALUATING INCOMPLETE MULTIPLE IMPERFECT DIAGNOSTIC TESTS WITH A PROBIT LATENT CLASS MODEL

Yi Zhang, University of North Carolina at Chapel Hill*
Haitao Chu, University of Minnesota
Donglin Zeng, University of North Carolina at Chapel Hill

Accurate diagnosis of a molecularly defined subtype of cancer is important toward its effective prevention and treatment. Since a gold standard may be unavailable, tumor sub-type status is commonly measured by multiple imperfect diagnostic markers. Furthermore, some subjects are only measured by a subset of diagnostic tests due to cost or compliance. In this paper, we propose a Probit latent class (PLC) model to model latent values for diagnostic tests within diseased and non-diseased groups. Unstructured correlations are used to model dependence among tests. EM algorithm is used to estimate diagnostic accuracy parameters, prevalence, and correlations. The proposed method is applied to analyze data from the NCI Colon Cancer Family Registry (C-CFR) on diagnosing microsatellite instability (MSI) for hereditary nonpolyposis colorectal cancer (HNPCC) with eleven biomarker tests. Simulations are conducted to evaluate the small-sample performance of our method.

e-mail: yzhang@bios.unc.edu

92. META-ANALYSIS

META-ANALYSIS OF BINARY RARE ADVERSE EVENT

Dulal K. Bhaumik, University of Illinois at Chicago
Anup K. Amatya, New Mexico State University*
Sharon-Lise Normand, Harvard University
Joel Greenhouse, Carnegie Mellon University
Eloise Kaizar, The Ohio State University
Brian Neelon, Duke University
Robert Gibbons, University of Chicago

We examine the use of fixed-effects and random-effects moment-based meta-analytic methods for analysis of binary adverse event data. Special attention is paid to the case of rare adverse events which are commonly encountered in routine practice. We study estimation of model parameters and between-study heterogeneity. In addition we examine traditional approaches to hypothesis testing of the average treatment effect and detection of the heterogeneity of treatment effect across studies. We then study the statistical properties of both the traditional and new methods via simulation. We find that in general, moment-based estimators of combined treatment effects and heterogeneity are biased and the degree of bias is proportional to the rarity of the event under study. The new methods eliminate much, but not all of this bias.

e-mail: aamatya@nmsu.edu

REGULATORY NETWORK ANALYSIS BY META-ANALYSIS OF MULTIPLE TRANSCRIPTOMIC STUDIES IN MAJOR DEPRESSIVE DISORDER

Ying Ding, University of Pittsburgh*
Etienne Sibille, University of Pittsburgh
George Tseng, University of Pittsburgh

Major depressive disorder (MDD) is a heterogeneous illness with mostly uncharacterized underlying genetics. Analysis of single MDD microarray study often faces difficulties from small sample size, weak signal and existence of confounding variables. In our previous analysis, we developed a random intercept model and meta-regression combining eight MDD microarray studies to identify ~420 disease associated candidate markers. In this current study, we will focus on the ~420 candidate genes for network construction using gene co-expression analysis by direct statistical correlation and liquid association analysis under a meta-analysis setting. We show that the combined network by meta-analysis is more accurate and stable than network analysis from single studies. The meta-network analysis provides a framework to effectively narrow down gene targets with driver gene capability or potential therapeutic targets.

e-mail: dingying85@gmail.com

META-ANALYSIS FRAMEWORK FOR THE DIMENSION REDUCTION OF GENOMIC DATA

Dongwan D. Kang, University of Pittsburgh*
George C. Tseng, University of Pittsburgh

Principal component analysis (PCA) enables researchers to explore high dimensional data through projection to a low-dimensional space. As an exploratory tool to visualize subjects in 2 or 3 dimensional subspace while minimizing information loss, PCA is one of the most popular multivariate analysis techniques. In this paper, we consider simultaneous dimension reduction using PCA when multiple genomic studies are combined. Although similar concepts of common principal components analysis exist, the advantage of such a practice in the meta-analysis context has not been studied. We propose two meta-analysis approaches to find a common principal component (PC) subspace by variance sum maximization and angle sum minimization criteria. We further extend the concept to incorporate robust PCA and sparse PCA in the meta-analysis framework. We evaluated the advantages and limitations of the proposed methods in the context of dimension reduction for data visualization and supervised machine learning using five examples of real genomic data.

e-mail: dok11@pitt.edu

META-ANALYSIS OF OBSERVATIONAL STUDIES WITH UNMEASURED CONFOUNDERS

Lawrence C. McCandless, Simon Fraser University, Canada*

Meta-analysis is a statistical method that is used to combine the results of different studies in order draw conclusions about a body of research. For example, one might imagine extracting hazard ratios and odds ratio from a collection of different health research papers looking at the effectiveness and safety of a drug (e.g. antidepressants). An emerging area of innovation in statistics involves meta-analysis of observational studies. Unlike randomized controlled trials, which are the gold standard for proving causation, observational studies are prone to biases such as confounding and measurement error. In this talk, I will present a novel methodology for meta-analysis of observational studies with unmeasured confounders. I draw parallels with sensitivity analysis and Bayesian inference. The discussion is motivated from a meta-analysis of the relationship between oral contraceptives and endometriosis.

e-mail: lmccandl@sfu.ca

COMPREHENSIVE COMPARATIVE STUDY OF MICROARRAY META-ANALYSIS METHODS

Lun-Ching Chang, University of Pittsburgh*
Hui-Min Lin, University of Pittsburgh
George C. Tseng, University of Pittsburgh

With the rapid data generation from microarray experiments in the past decade, genomic meta-analysis to combine information across multiple studies of relevant biological hypotheses has gained popularity. Many microarray meta-analysis methods have been developed and applied in the literature, of which different assumptions and biological goals are presumed. In the literature, the methods have only been minimally investigated and compared. In this talk, we will present a systematic comparison of twelve microarray meta-analysis methods in four large scale examples combining seven prostate cancer studies, seven brain cancer studies, eight major depressive disorder studies and six lung cancer studies, respectively. We will evaluate based on quantitative criteria of sensitivity, stability, robustness and biological association. The aggregated results will provide a practical guideline to the best choice of method(s) for a given application.

e-mail: lunching@gmail.com

IMPUTATION OF TRUNCATED p-VALUES FOR META-ANALYSIS METHODS AND ITS GENOMIC

Shaowu Tang, University of Pittsburgh*
George C. Tseng, University of Pittsburgh

Microarray analysis to monitor expression activities in thousands of genes simultaneously has become a routine experiment in biomedical research during the past decade and information integration by meta-analysis to detect differentially expressed (DE) genes has become popular to obtain increased statistical power and validated findings. In practice, the detected DE gene lists under certain p-value threshold (e.g. DE genes with p-value<0.001) are often reported in the journal publications. In order to avoid applying less efficient vote counting method or naively drop the studies with incomplete information, we develop effective meta-analysis methods for such situation with partially censored p-values. We developed and compared three imputation methods -- mean imputation, single random imputation and multiple imputation -- for a general class of evidence aggregation methods of which Fisher, Stouffer and logit methods are special examples. The null distribution of each method was analytically derived and subsequent inference and genomic analysis framework were established. Simulations were performed and the methods were applied to two genomic applications in prostate cancer and major depressive disorder. The results showed that imputation methods outperformed naive approaches.

e-mail: sht41@pitt.edu

MERGING CLUSTERED OR LONGITUDINAL COHORT DATA WITH COHORT-SPECIFIC MISSING COVARIATES

Fei Wang, University of Michigan*
Lu Wang, University of Michigan
Peter X.-K. Song, University of Michigan

Analyzing multiple datasets collected from similar cohort studies is often undertaken in practice. Here, we aim to develop statistical approaches that enable us to merge clustered or longitudinal datasets with strong heterogeneity, such as different within-subject correlations and follow-up schedules. Moreover, some covariates may be observed in some studies but completely unmeasured in the other studies. We propose an estimating equation approach to analyze these datasets jointly, in which we model the mechanism of missing covariates nonparametrically. Under some mild regularity conditions, we show the proposed estimator is consistent and asymptotically normal. Through simulation studies, our method has shown desirable finite sample performances and thus is recommended as a valid approach to merging clustered and longitudinal cohort datasets when some covariates are missing in certain cohorts.

e-mail: wafei@umich.edu

93. MISSING DATA I

A MULTIPLE IMPUTATION BASED APPROACH TO SENSITIVITY ANALYSES AND EFFECTIVENESS ASSESSMENTS IN LONGITUDINAL CLINICAL TRIALS

Teshome Birhanu, I-BioStat, Universiteit Hasselt, Belgium*
Ilya Lipkovich, Eli Lilly & Company
Geert Molenberghs, I-BioStat, Universiteit Hasselt, Belgium and I-BioStat, Katholieke Universiteit Leuven, Belgium
Craig H. Mallinckrodt, Eli Lilly & Company

It is important to understand the effects of a drug as actually taken (effectiveness) and when taken as directed (efficacy). The statistical performance of a method referred to as placebo multiple imputation (pMI) as an estimator of effectiveness and as a worst reasonable case sensitivity analysis in assessing efficacy was investigated. The pMI method assumes the statistical behavior of drug-treated patients after drop out is the statistical behavior of placebo-treated patients. Thus, in the effectiveness context pMI assumes no pharmacological benefit of the drug after dropout. In the efficacy context pMI is a specific form of an MNAR analysis expected to yield a conservative estimate of efficacy. In a simulation study with 18 scenarios the pMI approach generally provided

unbiased estimates of effectiveness and conservative estimates of efficacy. In contrast, LOCF and BOCF were conservative in some scenarios and anti-conservative in others with respect to efficacy and effectiveness. As expected, DL and MI yielded unbiased estimates of efficacy and tended to over-estimate effectiveness in those scenarios where a drug effect existed. However, in scenarios with no drug effect, and therefore the true values for both efficacy and effectiveness were zero, DL and MI yielded unbiased estimates of efficacy and effectiveness.

e-mail: birhanu.teshomeayele@uhasselt.be

ESTIMATION OF RATE OF CHANGE IN LONGITUDINAL STUDIES WITH VARYING DEGREES OF MISSINGNESS AND INFORMATIVE DROPOUT: A SIMULATION STUDY

Jamie E. Collins, Boston University*
Robin Bliss, Brigham and Women's Hospital
Elena Losina, Brigham and Women's Hospital

Informative dropout in longitudinal studies can lead to biased and inaccurate estimators when the dropout process is ignored. Many methods have been proposed to address this problem; however, each method comes with its own set of complex modeling assumptions. The performance of simple methods for longitudinal data that ignore informative dropout when overall dropout is relatively small or when only a portion of dropout is informative has not been examined. The goal of this study was to evaluate general linear mixed models that ignore informative dropout when estimating rate of change in longitudinal studies. Using a simulation study we compared the bias, accuracy, and coverage of the model under a wide range of scenarios including variations in the amount of overall dropout, the amount of informative dropout, and a range of sample sizes and standard deviations of change. When overall dropout was small (less than 15% by study end), mixed effects models that ignored the dropout mechanism tended to perform well, even when the missing data mechanism was informative. When dropout was severe (greater than 60% by study end), even small amounts of informative dropout led to substantial increased bias and decreased accuracy.

e-mail: collinsj@bu.edu

ON CLUSTER SIZE, IGNORABILITY, ANCILLARITY, COMPLETENESS, SEPARABILITY, AND DEGENERACY: SEQUENTIAL TRIALS, RANDOM SAMPLE SIZES, AND MISSING DATA

Geert Molenberghs, I-BioStat, Universiteit Hasselt & Katholieke Universiteit Leuven, Belgium*
Michael G. Kenward, London School of Hygiene and Tropical Medicine
Marc Aerts, I-BioStat, Universiteit Hasselt, Belgium
Geert Verbeke, I-BioStat, Katholieke Universiteit Leuven & Universiteit Hasselt, Belgium
Anastasios A. Tsiatis, North Carolina State University
Marie Davidian, North Carolina State University
Dimitris Rizopoulos, Erasmus University Rotterdam

Many statistical properties are derived for fixed sample size. Familiar results then follow, e.g., consistency, asymptotic normality, and efficiency of the sample average for the mean parameter. Matters change when sample size itself becomes random, either in a deterministic or probabilistic way, either data-dependent or not. Settings include sequential trials, missing data, and completely random sample size. While a lot of work has been done in this area, it is insightful to place this into a general joint-modeling framework. Then, parametric and semi-parametric inferences can be drawn. It is shown that counterintuitive results follow, e.g., the fact that the sample average may exhibit small-sample bias and, even when unbiased, like with a completely random sample size, then it is not optimal, without a uniform optimum existing. We demonstrate that such results critically depend on key attributes, such as (non-) ancillarity of the sample size and the fact that the sample sum combined with the sample size never is a so-called complete sufficient statistic. Our results have implications for estimation after group sequential trials. There are ramifications for other settings, such as random cluster sizes, censored time-to-event data, and joint modeling of longitudinal and time-to-event data.

e-mail: geert.molenberghs@uhasselt.be

DIAGNOSTIC PLOTS FOR EVALUATION OF BIAS IN MISSING DATA FROM CLINICAL TRIALS

Gerry W. Gray, U.S. Food and Drug Administration*

Often in a clinical trial there are substantial numbers of patients whose outcomes are missing. If the missing patients differ in important ways from the patients whose outcomes are observed then the outcome of the trial could be misleading. A simple plot, sometimes called a "tipping point" plot, of the potential results of the trial as a function of the outcomes of the missing patients can be informative. The addition of two additional layers of information, the value of the parameter of interest as compared to the value for the observed patients and the predictive distribution of the missing outcomes, can make this plot much more informative. These two additional layers also permit a visual representation of the sensitivity of the trial outcome to various departures from the MCAR assumption.

e-mail: gerry.gray@fda.hhs.gov

TIME-TO-EVENT ANALYSIS WITH PARTIAL ADJUDICATION OF POTENTIAL EVENTS USING FRACTIONAL IMPUTATION

Jason C. Legg, Amgen Inc.*
Jae Kwang Kim, Iowa State University

In clinical trials and large observational studies using databases, clinical events of interest are often not initially observed. Potential events such as a local investigator assessments, laboratory results, or treatment or procedure codes are recorded. For important events, identification of true events from the potential events could include using an adjudication committee, invasive or costly pro-

cedures, or medical chart reviews. These approaches add a cost per event to the study. In prospective studies, this translates into a random cost component, which complicates planning. For a large database in an observational study or for less critical events in a clinical trial, it can be infeasible to adjudicate all potential events. We propose using a measurement error subsample of potential events and parametric fractional imputation (Kim 2011) of events and nonevents. Estimators for functions of the underlying survival curves are proposed that incorporate unit weights from randomization or sampling. The approach builds on the data augmentation works of Snapinn (1988) and Cook and Kosorok (2004) in similar partial event information problems. Discussion on how to select the subsample is also presented.

e-mail: jlegg@amgen.com

MULTIPLE IMPUTATION FOR GENERALIZED LINEAR MODELS WITH CENSORED COVARIATES

Paul W. Bernhardt, North Carolina State University*
Huixia Wang, North Carolina State University
Daowen Zhang, North Carolina State University

Censored observations are a common occurrence in biomedical datasets. Though censoring is commonly associated with time-to-event data, censored data also arise due to detection limits. Very little research has focused on proper statistical procedures when predictors are censored due to detection limits. We propose a frequentist multiple imputation method for analyzing datasets with censored predictors within the context of generalized linear models. We establish the consistency and asymptotic normality of the proposed multiple imputation estimator and provide a consistent variance estimator. Through an extensive simulation study, we demonstrate that the proposed multiple imputation method leads to consistent parameter estimates while several competing estimators are biased, more variable, or computationally intensive to obtain. We apply the proposed multiple imputation method to analyze the GenIMS dataset which has several biomarkers subject to censoring due to lower detection limits.

e-mail: pwbernha@ncsu.edu

MULTIPLE IMPUTATION FOR MEASUREMENT ERROR WITH INTERNAL AND EXTERNAL CALIBRATION SAMPLES

Roderick J. Little, University of Michigan*

Existing methods for the analysis of data involving assay data subject to measurement error are deficient. In particular, classical calibration methods have been shown to yield invalid inferences unless the measurement error is small. Regression calibration, a form of conditional mean imputation, has better properties, but is not well suited to adjusting for heteroscedastic measurement error. Bayesian multiple imputation is less common for measurement error problems than for missing data, but we argue that it represents an attractive option in the measurement error, providing superior inferences to existing methods and a convenient way of adjusting for measurement error using simple complete-data methods and multiple imputation combining rules. It also provides a convenient approach to limit of quantification issues, another area where current approaches are in our view deficient. We review some recent work that develops multiple imputation methods for assay data, focusing particularly on three key aspects: internal versus external calibration designs, the role of the non-differential measurement error assumption in these designs, and heteroscedastic measurement error.

e-mail: rlittle@umich.edu

94. SEMIPARAMETRIC AND NONPARAMETRIC METHODS FOR SURVIVAL ANALYSIS

A FAMILY OF WEIGHTED GENERALIZED INVERSE WEIBULL DISTRIBUTION

Broderick O. Oluyede, Georgia Southern University*
Jing Kersey, East Georgia College

A family of weighted generalized inverse Weibull distribution called the beta-inverse Weibull distribution is proposed. We present theoretical properties of the distribution. We also discuss useful transformations that leads to generation of observations from the proposed distribution, as well as estimation of the parameters of the distribution.

e-mail: boluyede@georgiasouthern.edu

STRATIFIED AND UNSTRATIFIED LOG-RANK TESTS IN SURVIVAL ANALYSIS

Changyong Feng, University of Rochester Medical Center*
David Oakes, University of Rochester Medical Center
Yao Yu, University of Rochester Medical Center

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 closed form 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 study the robustness of the combined test by simulations.

e-mail: feng@bst.rochester.edu

NONPARAMETRIC ESTIMATION OF THE MEAN FUNCTION FOR RECURRENT EVENTS DATA WITH MISSING EVENT CATEGORY

Feng-Chang Lin, University of North Carolina at Chapel Hill*
Jianwei Cai, University of North Carolina at Chapel Hill
Jason P. Fine, University of North Carolina at Chapel Hill
HuiChuan J. Lai, University of Wisconsin-Madison

Recurrent event data frequently arise in longitudinal studies when study subjects possibly experience more than one event during the observation period. Often, such recurrent events can be categorized. An analysis that incorporates such categorization is more informative than the one that aggregates information across categories. However, part of the categorization may be missing due to technical difficulties or recording ignorance. When the researchers are interested in a mean function without specifying any parametric/semiparametric form, a complete-case analysis would underestimate the truth even when the event category is missing completely at random. In this research we study a nonparametric approach for the estimation of the mean function by utilizing local polynomial regression techniques to estimate the probability of the event category when the missing information is present. Consistency and large sample normality of our estimators are proved. Simulation results show that our estimation is close to the Nelson-Aalen estimator which requires no missingness on event category. The proposed method was applied to the cystic fibrosis (CF) registry data.

e-mail: flin@bios.unc.edu

MEDIAN TESTS FOR CENSORED SURVIVAL DATA: CONTINGENCY TABLE APPROACH

Shaowu Tang, University of Pittsburgh
Jong-Hyeon Jeong, University of Pittsburgh*

We modify a K-sample median test for censored survival data (Brookmeyer and Crowley, 1982) through a simple contingency table approach where each cell counts the number of observations in each sample that are greater than the pooled median or vice versa. Under censoring, this approach would generate non-integer entries for the cells in the contingency table. We propose to construct a weighted asymptotic test statistic that aggregates dependent chisquare statistics formed at the nearest integer points to the original non-integer entries. We show that this statistic follows approximately a chisquare distribution with k-1 degrees of freedom. For a small sample case, we propose a test statistic based on combined p-values from Fisher's exact tests, which follows a chisquare distribution with 2 degrees of freedom. Simulation studies are performed to show that the proposed method provides reasonable type I error probabilities and powers. The proposed method is illustrated with two real datasets from phase III breast cancer clinical trials.

e-mail: jeong@nsabp.pitt.edu

POINTWISE CONFIDENCE INTERVALS FOR A SURVIVAL DISTRIBUTION FOR RIGHT CENSORED DATA WITH SMALL SAMPLES OR HEAVY CENSORING

Michael P. Fay, National Institute of Allergy and Infectious Diseases, National Institutes of Health*
Erica Brittain, National Institute of Allergy and Infectious Diseases, National Institutes of Health
Michael A. Proschan, National Institute of Allergy and Infectious Diseases, National Institutes of Health

We propose a nonparametric confidence procedure for the survival function at any specified time for right-censored data assuming only independent censoring. In such situations, typically the Kaplan-Meier curve is used together with some kind of asymptotic confidence interval. These asymptotic confidence intervals can have coverage problems when the number of observed failures is not large, and/or when testing the latter parts of the curve where there are few remaining subjects at risk. Our procedure is designed to have good coverage in those situations. When there is no censoring, our procedure reduces to the exact binomial Clopper-Pearson confidence interval. We show that our procedure is asymptotically equivalent to using a confidence interval on the Kaplan-Meier curve using Greenwood's variance, and hence our procedure gives asymptotically correct coverage. The procedure may be inverted to create confidence procedures for a quantile (e.g., median) of the survival distribution. Simulations confirm that our procedure retains the type I error rate in many situations where competing methods do not.

e-mail: mfay@niaid.nih.gov

FURTHER THOUGHTS ON THE PROPORTIONAL MEAN RESIDUAL LIFE MODEL

David Oakes, University of Rochester Medical Center*

The proportional mean residual life model was introduced by Oakes and Dasu (1990) and has since been studied by a number of authors, including, notably, Chen and Cheng (2006). When the mean residual life functions are decreasing functions of time, the model is a special case of an excess risk model discussed by, amongst others, Sasieni (1996) and Martinussen and Scheike (2002). Some implications of this representation for inference about the model parameters are discussed.

e-mail: oakes@bst.rochester.edu

FRAILTY MODELS WITH COVARIATES SUBJECT TO LIMIT OF DETECTION

Abdus Sattar, Case Western Reserve University*
Liang Li, Cleveland Clinic Foundation
Pingfu Fu, Case Western Reserve University

In many biomedical and genetic epidemiological studies biomarkers are measured routinely. Sometimes the actual value of the biomarker measurement is unobserved but known to be below a threshold, called the limit of detection (LOD). This is a form of left censoring. Naïve approaches ignoring this problem, such as replacing the undetected value by LOD or half of the LOD, often produce bias and inefficient hazard ratio estimates in frailty models for time-to-event. We proposed and compared several alternative approaches to this problem in the context of a partially linear frailty model: multiple imputation, Monte Carlo EM, and inverse censoring probability weighting. We illustrate the use of the proposed methods with a data set from an HIV study in which viral load was subject to lower detection limit.

e-mail: sattar@case.edu

95. NEW STATISTICAL CHALLENGES IN FUNCTIONAL DATA ANALYSIS

BAYESIAN VARIABLE SELECTION FOR IDENTIFYING GENETIC EFFECTS ON FUNCTIONAL CONNECTIVITY

Brian J. Reich, North Carolina State University*
Michele Guindani, University of Texas MD Anderson Cancer Center
Abel Rodriguez, University of California at Santa Cruz
Vince Calhoun, University of New Mexico

Functional magnetic resonance imaging (fMRI) data are often used to identify regions of the brain that are functionally connected while performing a cognitive task. Connectivity patterns vary across subjects according to subject-specific characteristics, e.g., the subject having been diagnosed with a form of schizophrenia. Our objective is to identify genetic pathways that affect a subject's functional connectivity in response to a series of external stimuli. We model each subject's connectivity using a graphical model, with potentially a different set of edges for each subject. We assume that the probability of each pair of regions being connected depends on a set of subject-specific genetic covariates. This gives a high-dimensional model, as the number potential region pairs and the number of genetic variables are both large. Therefore, we propose a Bayesian variable selection technique to identify a sparse model for functional connectivity. The approach is illustrated on a a set of genetic and fMRI data from a population of healthy and schizophrenic patients.

e-mail: brian_reich@ncsu.edu

REGRESSION MODELS FOR SPATIALLY CORRELATED MULTILEVEL FUNCTIONAL DATA

Ana-Maria Staicu, North Carolina State University*
Damla Sentürk, University of California at Los Angeles
Raymond J. Carroll, Texas A&M University

Regression with a functional predictor is now a well-studied area in functional data analysis. Existing methods, however, are limited to scalar responses (continuous or discrete), when the predictor has, furthermore, a multilevel functional structure. The setting where both the response and predictor have more complex functional structure raises many computational and theoretical challenges. We introduce a time varying regression framework for the case when both the response and predictor are functional data with a hierarchical structure such that the functions at the lowest hierarchy level are spatially correlated. This work combines functional data and spatial statistics tools in order to propose parsimonious and computationally feasible functional regression models. The proposed approach is easily extendable to the setting where the functional response is discrete-valued. Our methods are inspired by and applied to data obtained from a state-of-the-art colon carcinogenesis scientific experiment, where of interest is to describe the relation between the apoptosis indicator and the concentration of the biomarker p27 measured at cellular level, for each of many colonic crypts within different rats, while accounting for the crypt signaling.

e-mail: ana-maria_staicu@ncsu.edu

VARYING COEFFICIENT MODELS FOR SPARSE NOISE-CONTAMINATED LONGITUDINAL DATA

Damla Senturk, University of California, Los Angeles*
Danh Nguyen, University of California, Davis

In this paper we propose a varying coefficient model for sparse longitudinal data that allows for error-prone time-dependent variables and time-invariant covariates. We develop a new estimation procedure, based on covariance representation techniques, that enables effective borrowing of information across all subjects in sparse and irregular longitudinal data observed with measurement error, a challenge for which there is no current adequate solution. Sparsity is addressed via a functional analysis approach that considers the observed longitudinal data as noise contaminated realizations of a random process that produces smooth trajectories. This approach allows for estimation based on pooled data, borrowing strength from all subjects, in targeting the mean functions and auto- and cross-covariances to overcome sparse noisy designs. The resulting estimators are shown to be uniformly consistent. Consistent prediction for the response trajectories are also obtained via conditional expectation under Gaussian assumptions. Asymptotic distributions of the predicted response trajecto-

ries are derived, allowing for construction of asymptotic pointwise confidence bands. Efficacy of the proposed method is investigated in simulation studies and compared to the commonly used local polynomial smoothing method. The proposed method is illustrated with a sparse longitudinal data set, examining the age-varying relationship between calcium absorption and dietary calcium.

e-mail: dsenturk@ucla.edu

LONGITUDINAL HIGH DIMENSIONAL DATA ANALYSIS

Vadim Zipunnikov, Johns Hopkins University*
Sonja Greven, Ludwig-Maximilians-University
Brian Caffo, Johns Hopkins University
Daniel S. Reich, Johns Hopkins University and National Institute of Neurological Disorders and Stroke, National Institutes of Health
Ciprian M. Crainiceanu, Johns Hopkins University

We develop a flexible framework for modeling high-dimensional functional and imaging data observed longitudinally. The approach decomposes the observed variability of high-dimensional observations measured at multiple visits into three additive components: a subject-specific functional random intercept that quantifies the cross-sectional variability, a subject-specific functional slope that quantifies the dynamic irreversible deformation over multiple visits, and a subject-visit specific functional deviation that quantifies exchangeable or reversible visit-to-visit changes. The proposed method is very fast, scalable to studies including ultra-high dimensional data, and can easily be adapted to and executed on modest computing infrastructures. The method is applied to the longitudinal analysis of diffusion tensor imaging (DTI) data of the corpus callosum of multiple sclerosis (MS) subjects. The study includes 176 subjects observed at 466 visits. For each subject and visit the study contains a registered DTI scan of the corpus callosum at roughly 30,000 voxels.

e-mail: vzipunni@jhsph.edu

96. ESTIMATION OF COVARIANCE MATRICES WITH APPLICATIONS TO LONGITUDINAL DATA AND GRAPHICAL MODELS

ESTIMATING LARGE CORRELATION MATRICES BY BANDING THE PARTIAL AUTOCORRELATION MATRIX

Yanpin Wang, University of Florida
Michael Daniels, University of Florida*

In this article, we propose a computationally efficient approach to estimate (large) p-dimensional correlation matrices of ordered data based on an independent sample of size n. To do this, we construct the estimator based on a k-band partial autocorrelation matrix with the number of bands chosen using an exact multiple hypothesis testing procedure. This approach is considerably faster than many existing methods and only requires inversion of k dimensional covariances matrices. In addition, the resulting estimator is guaranteed to be positive definite as long as $k < n$ (even when $n < p$). We evaluate our estimator via extensive simulations and compare it to the Ledoit-Wolf estimator. We also illustrate the approach using high-dimensional sonar data.

e-mail: mdaniels@stat.ufl.edu

ANTEDEPENDENCE MODELS FOR NORMAL AND CATEGORICAL LONGITUDINAL DATA

Dale L. Zimmerman, University of Iowa*

Antedependence models are useful generalizations of well-known stationary autoregressive models or Markov models for longitudinal data, which allow for the strength and order of serial dependence to change over time. Many likelihood-based inferential methods for such models are of closed-form or are computationally very easy, even when data are (ignorably) missing. In this presentation I describe some of these methods (including maximum likelihood estimators and likelihood ratio tests for order of antedependence) for normal and categorical longitudinal data. Their usefulness at revealing and accommodating important features of serial dependence are illustrated with two examples.

e-mail: dale-zimmerman@uiowa.edu

DOUBLY REGULARIZED ESTIMATION AND SELECTION IN LINEAR MIXED-EFFECTS MODELS FOR HIGH-DIMENSIONAL LONGITUDINAL DATA

Yun Li, University of Michigan
Sijian Wang, University of Wisconsin
Peter X.K. Song, University of Michigan
Naisyin Wang, University of Michigan
Ji Zhu, University of Michigan*

The linear mixed effects model (LMM) is widely used in the analysis of clustered or longitudinal data. This paper aims to address analytic challenges arising from parameter estimation and effects selection in the application of the LMM for high-dimensional longitudinal data. We develop a doubly regularized approach in the LMM to simultaneously select fixed and random effects. On the theoretical front, we establish large sample properties for the proposed method when both numbers of fixed effects and random effects diverge to infinity along with the sample size. We present regularity conditions for the diverging rates, under which the proposed method achieves both estimation and selection consistency. In addition, we propose a new algorithm that solves the related optimization problem effectively so that its computational cost is comparable with that of the Newton-Raphson algorithm for maximum likelihood estimator in the LMM. Simulation studies and data examples are also used to demonstrate the performance of the proposed method.

e-mail: jizhu@umich.edu

97. ANALYSES OF INCOMPLETE LONGITUDINAL DATA –HOW ROBUST ARE THE RESULTS?

BAYESIAN INFLUENCE MEASURES FOR JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA

Joseph G. Ibrahim, University of North Carolina at Chapel Hill*
Hongtu Zhu, University of North Carolina at Chapel Hill
Niansheng Tang, Yunnan University, China

We develop a variety of influence measures for carrying out perturbation (or sensitivity) analysis to joint models of longitudinal and survival data (JMLS) in Bayesian analysis. A perturbation model and its associated perturbation manifold are introduced to characterize individual and global perturbations to the three components of a Bayesian model, including the data points, the prior distribution and the sampling distribution. Local influence measures are proposed to quantify the degree of these perturbations to JMLS. The proposed methods allow the detection of outliers or influential observations and the assessment of the sensitivity of inferences to various unverifiable assumptions on the Bayesian analysis of JMLS. Simulation studies and a real dataset are used to highlight the broad spectrum of applications for our Bayesian influence methods.

e-mail: ibrahim@bios.unc.edu

ROBUST ANALYSES OF RANDOMIZED CLINICAL TRIALS WITH INCOMPLETE LONGITUDINAL DATA

Devan V. Mehrotra, Merck Research Laboratories*

In a typical randomized clinical trial, a continuous variable of interest is measured at baseline and fixed post-baseline time points. The resulting longitudinal data, often incomplete due to dropouts, are commonly analyzed using parametric likelihood-based methods that assume multivariate normality of the response vector. If the normality assumption is deemed untenable, then semi-parametric methods such as (weighted) generalized estimating equations are considered. We propose an alternate approach in which the missing data problem is tackled using multiple imputation, and each imputed dataset is analyzed using robust regression (M-estimation) to protect against potential non-normality/outliers. The results from each imputed dataset are combined for overall inference using either the simple Rubin (1987) method, or the more complex but potentially more accurate Robins and Wang (2000) method. We use simulations to show that our proposed approach performs at least as well as the standard methods under normality, but is notably better under both elliptically symmetric and asymmetric non-normal distributions. A real clinical trial is used for illustration.

e-mail: devan_mehrotra@merck.com

ON THE USEFULNESS OF SENSITIVITY ANALYSES

James M. Robins, Harvard School of Public Health*

I review methods for sensitivity analyses of longitudinal data with time dependent treatments and informative censoring. I follow with a critical discussion of the usefulness of these analyses for medical decision making.

e-mail: robins@hsph.harvard.edu

98. Statistics in Mental Health Research: A Memorial to Dr. Andrew Leon

Dr. Andrew Leon, originally scheduled to participate in this session, passed away unexpectedly on Sunday, February 19, 2012. The remaining participants would like to dedicate their session to his memory.

EFFICIENT LONGITUDINAL ESTIMATION OF INCIDENCE AND PREVALENCE RATE OF MAJOR DEPRESSIVE DISORDER IN HOME HEALTHCARE STUDY

Samiran Ghosh, Winthrop University Hospital and SUNY Stony Brook, Marty Bruce, Weill Cornell Medical College*

sghosh@winthrop.org

MODELING BETWEEN- AND WITHIN-SUBJECT MOOD VARIANCE IN ECOLOGICAL MOMENTARY ASSESSMENT (EMA) DATA USING MIXED-EFFECTS LOCATION-SCALE MODELS

Donald Hedeker, University of Illinois at Chicago*
Robin J. Mermelstein, University of Illinois at Chicago
Hakan Demirtas, University of Illinois at Chicago

Modern data collection procedures, such as ecological momentary assessments (EMA), experience sampling, and diary methods, have been developed to record the momentary events and experiences of subjects in their daily lives. In EMA studies, it is common to have dozens or more observations per subject, allowing greater opportunities for within-subject modeling. In particular, one very promising approach is the modeling of both between-subject (BS) and within-subject (WS) variances as a function of covariates, in addition to their effect on overall mean levels. In this presentation, we present data from an adolescent study in which subjects carry a palm pilot for a week and respond to random prompts and event-triggered prompts (when they smoke). We describe how mixed models, including heterogeneous between- and within-

subjects variance with random subject scale effects, can be used to address key mental health research questions. For example, we can investigate the degree to which changes in smoking across time are associated with changes in mood variation attributable to smoking. Also, by including the random scale effect we estimate the heterogeneity in the within-subjects variance. Thus, our approach allows us to investigate an important issue in mental health research, such as the effect of smoking on mood levels/variation as a person progresses in his/her smoking "career."

e-mail: hedeker@uic.edu

ARE ANTIDEPRESSANTS EFFECTIVE AND DO THEY CAUSE SUICIDAL THOUGHTS AND BEHAVIOR? METHODOLOGY AND FINDINGS FOR SYNTHESIZING FINDINGS ACROSS MULTIPLE RANDOMIZED ANTIDEPRESSANT TRIALS

Hendricks Brown, University of Miami*
Robert D. Gibbons, University of Chicago
Kwan Hur, University of Chicago and Hines VA Hospital Center for Medication Safety
John J. Mann, Columbia University
Bengt O. Muthen, University of California at Los Angeles

In 2004 and 2006 the FDA instituted a black-box warning on antidepressants in youth and young adults because of 1) data that suggested more suicidal ideation and behavior in youth given antidepressants compared to youth given placebo, and 2) very few individual trials in youth have shown significant reduction in depressive symptoms. Since these policy decisions, we have assembled and analyzed individual level data from 41 placebo controlled RCTs and conducted analyses to examine 1) overall impact of antidepressants on the course of depressive symptoms and suicidal ideation and behavior, 2) variation in impact as a function of age, sex and baseline depressive severity, 3) mediation effects of depressive symptoms on suicide ideation and behavior, and 4) causal inference. The statistical methods for such syntheses rely on new techniques for integrative data analysis (IDA) that combine individual level data from the different trials. These models incorporate multilevel growth and growth mixture modeling with latent variable and mediational modeling and rely on methods that have been developed by Gibbons et al., (under review), Brown et al., (2011 Prevention Science), and Muthén & Brown (2010 Statistics in Medicine) in the analysis of these data. Methods, results, and implications for other data synthesis studies are discussed.

e-mail: chbrown@med.miami.edu

THE FUTURE OF MENTAL HEALTH MEASUREMENT

Robert D. Gibbons, University of Chicago*

Mental health measurement has been based primarily on subjective judgment and classical test theory. Impairment level is usually determined by a total score, requiring that all respondents be administered the same items. An alternative to full scale administration is adaptive testing, in which different individuals may receive different scale items that are targeted to their specific impairment levels. This approach to testing is implemented via computerized adaptive testing (CAT) and multidimensional item response theory and is immediately applicable to a wide range of psychiatric measurement problems. We have developed a CAT depression inventory (CAT-DI) that can be administered adaptively, such that each individual responds only to those items that are most appropriate to assessing his/her level of depression. From a bank of 389 depression items, we can measure depression with an average of 12 items per subject and maintain a correlation of r=0.95 with the total 389 depression score. Applications to large scale screening for psychiatric epidemiology, genetic phenotyping, screening for depression in primary care, longitudinal measurement, and differential item functioning across different populations (e.g., Hispanic) are discussed.

e-mail: rdg@uchicago.edu

99. HIGH-IMPACT STATISTICAL METHODS AND THE FIGHT AGAINST HIV IN THE DEVELOPING WORLD

USING AUXILIARY BIOMARKERS TO IMPROVE POOLING STRATEGIES FOR HIV VIRAL LOAD TESTING

Tao Liu, Brown University*
Joseph W. Hogan, Brown University
Shangxuan Zhang, Brown University
Rami Kantor, Brown University

Monitoring HIV viral load (VL) is critical for infected individuals taking antiretroviral medications. Typically the goal is to determine whether the VL exceeds a certain threshold, which indicates treatment failure. However, in resource limited settings, individual VL testing is prohibitively expensive. The use of sample pooling prior to VL testing can reduce overall cost when the prevalence of those with detectable VL is low (e.g. less than 20%). We show how the use of routinely available and lower-cost biomarker data, such as CD4 count, can further reduce overall cost of treatment monitoring. The strategies are calibrated on a cohort study where VL is available on all individuals; potential cost savings are demonstrated using microsimulation.

e-mail: tliu@stat.brown.edu

THE ROLE OF NETWORK ANALYSES IN RESEARCH ON PREVENTION OF HIV INFECTION

Ravi Goyal, Harvard University
Joseph Blitzstein, Harvard University
Victor DeGruttola, Harvard University*

Efforts at prevention of HIV can be aided by an understanding of the transmission networks along which infection spreads in several ways: (1) characterization of the conditions under which interventions can succeed in controlling the HIV epidemic, (2) identification of appropriate means of tailoring implementation strategies to local conditions, and (3) determination of the level of adherence with community-wide interventions needed for successful control. However it is challenging to identify the most valuable network features and to obtain sufficiently reliable estimates of them. Here we consider estimation of network features from a sample, focussing on estimating the degree-degree mixing matrix, which quantifies assortativity, and discuss its use for network construction. Such construction is valuable in the development of epidemic models that can be used for testing potential value of intervention strategies though simulation. The methods will be investigated using a data set characterizing the sexual network in Likoma Island, Malawi. We also consider ways in which network-level information can be incorporated into cluster randomized trials to improve efficiency, and interpretability of results.

e-mail: degrut@hsph.harvard.edu

ESTIMATION FROM DOUBLE-SAMPLED SEMI-COMPETING RISK DATA

Constantin T. Yiannoutsos, Indiana University School of Medicine*
Menggang Yu, Indiana University School of Medicine
Hai Liu, Indiana University School of Medicine

In semi-competing risk data, a terminal event censors a non-terminal event, but not vice versa. We consider a situation when obtaining information about the terminal event from all subjects is not possible. An estimation procedure based on copula models (Fine, Jiang & Chappel, Biometrika, 2001, Lakhal, Rivest & Abdous, Biometrics, 2008) is adjusted for the incomplete ascertainment of the terminal event through double sampling of a random sample of subjects who have dropped out. Information about the terminal event is is ascertained only on these (double sampled) subjects. The performance of the proposed method is demonstrated via asymptotic study, simulations, and analysis of data from a large care and treatment program in sub-Saharan Africa where, because of large numbers of patient dropouts, vital status (and thus death, the terminal event) is ascertained on only a random sample of all subjects.

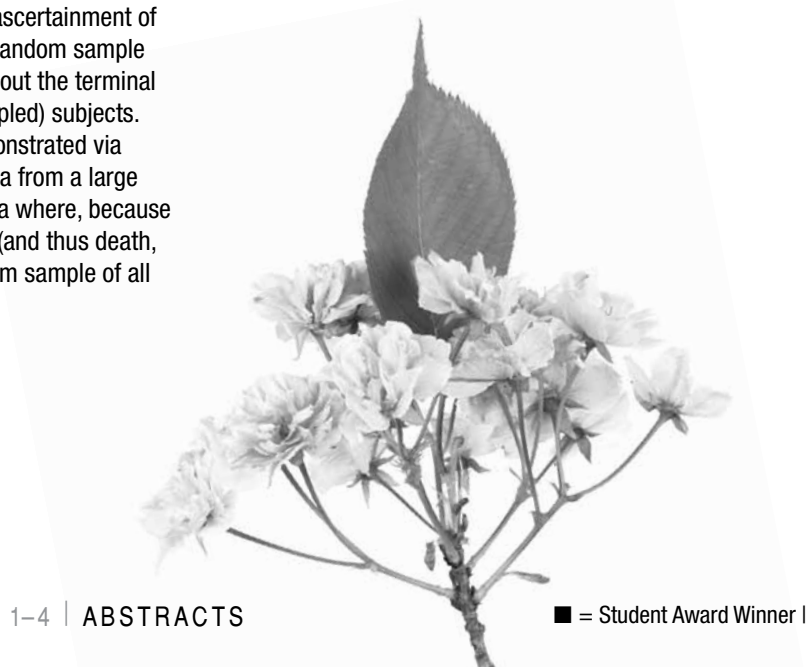
e-mail: cyiannou@iupui.edu

TRADITIONAL AND 'CAUSAL' MODELS FOR EVALUATING THE EFFECTIVENESS OF THE SWITCH TO SECOND LINE THERAPY IN A LARGE, ONGOING HIV/AIDS TREATMENT AND CARE PROGRAM IN A RESOURCE LIMITED SETTING

Sehee Kim, Harvard School of Public Health
Donna Spiegelman, Harvard School of Public Health*
Claudia Hawkins, Northwestern University
Aisa Muya, Management and Development for Health Dar es Salaam, Tanzania
Eric Aris, Management and Development for Health Dar es Salaam, Tanzania
Ester Mungure, Harvard School of Public Health
Aveika Akum, Harvard School of Public Health
Guerino Chalamilla, Management and Development for Health Dar es Salaam, Tanzania
Wafaie W. Fawzi, Harvard School of Public Health

Between 2004 and 2010, 94,598 patients were enrolled in HIV/AIDs Care and Treatment clinics in Dar es Salaam, Tanzania, with the support of the PEPFAR program, in partnership with the Harvard School of Public Health. Of these, 42,160 met the Tanzanian government's eligibility criteria for initiation of first line antiretroviral therapy (ARV1) and had sufficient data to be included in analysis, and of those, 2138 met the Tanzanian government's eligibility criteria for switching to second line ARVs (ARV2). 821 were switched to ARV2s, of whom 25% met the formal eligibility criteria at the time they were switched. The goal of this analysis is to assess the effectiveness and cost-effectiveness of 2ARVs in this large-scale community-based setting, by comparing outcomes among those who were switched to those who were eligible for switching but were not switched. The Kaplan-Meier estimate of the 6-month mortality among switchers was half that of eligible non-switchers, and the 1 year mortality was 30% lower. We will adjust these findings for confounding by indication, time-varying confounding, and immortal person-time bias, accounting for the lag between eligibility for switching and second line initiation (median 7.2 months). Results will be presented from several analytic options using standard and newer causal methods.

e-mail: stdls@hsph.harvard.edu



■ = Student Award Winner | * = Presenter

100. MEMORIAL SESSION FOR TOM TEN HAVE

CELEBRATING THE LIFE OF THOMAS R. TEN HAVE

J. Richard Landis, University of Pennsylvania*

This talk will provide highlights of Tom's life and career, touching on his many spheres of influence as family, friend, colleague, advisor, mentor, scientist, scholar, champion for social justice and humble servant. He tackled complex, socially urgent problems, contributing both to design and analysis innovations in personalized treatment strategies. He had a dream that the world should, and could, be better than the present realities of injustice. No one combined his professional pursuits w/ championing social justice more effectively than Tom. He was persistent in championing opportunities for underrepresented minorities at all levels. Even more than Tom's far-reaching professional impact, the way he treated each person he met was absolutely exemplary. He was always focused on others, even when struggling w/ his own battle for health. Tom's scholarship, integrity, and humble service continues to inspire all who knew him to reach out beyond themselves to encourage and invest in others.

e-mail: jrlandis@upenn.edu

SIZING SEQUENTIAL, MULTIPLE ASSIGNMENT, RANDOMIZED TRIALS FOR SURVIVAL ANALYSIS

Zhiguo Li, Duke University
Susan Murphy, University of Michigan*

Sequential Multiple Assignment Randomized Trials are growing in importance in developing dynamic treatment regimes. Usually the first stage involves randomization to one of several initial treatments. The second stage of treatment begins when an early nonresponse criterion or response criterion is met. In the second stage nonresponding subjects are rerandomized among second-stage treatments. Sample size calculations for planning these two-stage randomized trials with failure time outcomes are challenging because the variances of common test statistics depend in a complex manner on the joint distribution of time to the early nonresponse criterion or response criterion and the primary failure time outcome. We describe simple, albeit conservative, sample size formulae by using upper bounds on the variances. The resulting formulae only require the same working assumptions needed to size a standard single stage randomized trial and, in common settings are only mildly conservative. These sample size formulae are based on either a weighted Kaplan-Meier estimator of survival probabilities at a fixed time point or a weighted version of the log rank test.

e-mail: samurphy@umich.edu

POST-RANDOMIZATION MODIFICATION OF INTENT-TO-TREAT EFFECTS IN RANDOMIZED CLINICAL TRIALS

Rongmei Zhang, U.S. Food and Drug Administration*
Marshall Joffe, University of Pennsylvania
Thomas Ten Have, University of Pennsylvania

In the field of behavioral science, investigations involve the estimation of the effects of behavioral interventions on final outcomes for individuals stratified by post-randomization moderators measured during the early stages of the intervention. Motivated by this, we discuss the use of standard and causal approaches to assessing the modification of intent-to-treat effects of a randomized intervention by a post-randomization factor. We show analytically the bias of the estimators of the standard regression model under different combinations of assumptions. Such results show that the assumption of independence between two factors involved in an interaction, which has been assumed in the literature, is not necessary for unbiased estimation. Then, we present a structural nested distribution model estimated with G-estimation equations, which does not assume that the post-randomization variable is effectively randomized to individuals. Optimal G-estimator is also derived to achieve efficiency. Finally, we conduct simulations to compare the performance of causal and standard approaches and further assess our approach with data from a randomized cognitive therapy trial.

e-mail: rongmeiz@gmail.com

MEDIATION ANALYSES ON THE BASIS OF INITIAL RANDOMIZATION

Marshall Joffe, University of Pennsylvania*

Most methods for mediation analysis are based on assumptions of sequential ignorability. In randomized trials with a randomized intervention and an intermediate mediator, this means that not only is the initial treatment randomized but that effectively the mediator is as well (after possibly accounting for confounding variables). In randomized trials, the initial ignorability is guaranteed by randomization, but subsequent treatment decisions are not guaranteed to be ignorable. Mediation analyses can be performed of initial randomization which circumvent the need to assume implausible ignorability assumptions but come at the price of increased model dependence and decreased efficiency. This talk will highlight some of the contributions of Thomas Ten Have to the development, evaluation, dissemination, and application of these methods.

e-mail: mjoffe@mail.med.upenn.edu

101. ADVANCED STATISTICAL MODELING FOR COMPLEX OMICS DATA

BAYESIAN MODEL FOR IDENTIFYING SPATIAL INTERACTIONS OF CHROMATINS

Shili Lin, The Ohio State University*
Liang Niu, The Ohio State University

The expression of a gene is usually controlled by the regulatory elements in its promoter region. However, it has long been hypothesized that, in complex genomes, such as the human genome, a gene may be controlled by distant enhancers and repressors. A recent molecular technique, 3C (chromosome conformation capture), that uses formaldehyde cross-linking and locus-specific PCR was able to detect physical contacts between distant genomic loci. Such communication is achieved through spatial organization (looping) of chromosomes to bring genes and their regulatory elements into close proximity. Several adaptations of the 3C assay to study genome-wide spatial interactions, including Hi-C and ChIA-PET (chromatin interaction analysis by pair-end tag sequencing), have been proposed. However, due to the enrichment of ligation products on beads, such methods may also detect random collisions in addition to true spatial interactions. In this talk, I will present a hierarchical Bayesian model for analyzing such large-scale genome-wide looping data. In particular, I'll discuss how to use part of the correlated data to tackle the problem of random collision to reduce false positives. Data on DNA-protein binding and gene expression will also be integrated into the analysis to further enhance true loop detection power.

e-mail: shili@stat.osu.edu

TESTING AND ESTIMATION OF PARTIAL CORRELATION NETWORKS

Fred A. Wright, University of North Carolina at Chapel Hill*
Min Jin Ha, University of North Carolina at Chapel Hill

We describe a statistical framework to estimating partial correlation networks, while simultaneously performing valid hypothesis testing for network edges. We perform simple modifications of sample correlation matrices to handle singularities in estimating partial correlations, utilizing a “best of breed” approach to identify the best among a series of estimation procedures applied to a particular dataset. Challenges in estimation of standard errors and construction of p-values are handled using resampling approaches. Extensive simulation studies demonstrate the performance of various estimation methods under various scenarios, enabling methods with high accuracy to be chosen for each dataset. Our graph construction procedure is applied to yeast cell cycle microarray expression data, and the results validated using known pathways.

e-mail: fred_wright@unc.edu

STATISTICAL METHODS FOR INFERENCE FROM MULTIPLE ChIP-Seq SAMPLES

Sunduz Keles, University of Wisconsin, Madison*

As ChIP-seq technology is becoming more economical, generation of multiple ChIP-seq samples to elucidate contribution of transcription factor binding and epigenome to phenotypic variation is becoming standard. ChIP samples collected from different tissue types and/or individuals enable characterization of systematic changes in transcription factor binding and epigenomic patterns during development (intra-individual) or at the population level (inter-individual). Current analytical approaches for the analysis of ChIP-seq data are geared towards single sample investigations, and therefore have limited applicability in these comparative settings. We address this limitation by developing probabilistic models tailored for a rich class of multiple sample ChIP-seq problems.

e-mail: keles@stat.wisc.edu

STATISTICAL MODELS FOR ANALYZING SEQUENCING APPLICATIONS

Zhaohui S. Qin, Emory University*

The recent arrival of ultra-high throughput, next generation sequencing (NGS) technologies has revolutionized the genetics and genomics fields by allowing rapidly and inexpensively sequencing of billions of bases. The rapid deployment of NGS in a variety of sequencing-based experiments has resulted in fast accumulation of massive amount of sequencing data. To process this new type of data, a torrent of increasingly sophisticated algorithms and software tools are emerging to help the analysis stage of the NGS applications. Here we strive to comprehensively identify the critical challenges that arise from all stages of NGS data analysis and provide an objective overview of what have been achieved in existing works. At the same time, we highlight selected areas that need much further research to improve our current capabilities to delineate the most information from NGS data.

e-mail: zhaohui.qin@emory.edu

A GENE-TRAIT SIMILARITY REGRESSION METHOD FOR COMMON AND RARE VARIANTS WITH GENERAL TRAIT VALUES

Jung-Ying Tzeng, North Carolina State University*

We introduce a gene-trait similarity model to aggregate information from loci that are in the same gene or exonic region to study genetic effects . The method uses genetic similarity to aggregate information from multiple polymorphic sites, with adaptive weights dependent on allele frequencies and functionality scores to signify rare and common functional variants. Collapsing information at the similarity level instead of the genotype level avoids canceling signals with opposite etiological effects, is applicable to any class

of genetic variants without having to dichotomize the allele types, and can capture non-additive effects among markers by using certain similarity metrics. To assess gene-trait associations, trait similarities for pairs of individuals are regressed on their genetic similarities, with a score test whose limiting distribution is derived. We show how this framework can be applied to various of trait types such as continuous, binary, and survival traits.

e-mail: jytzeng@stat.ncsu.edu

102. BIOMARKERS II

THE APPLICATION OF NON-LINEAR MODELS TO UNDERSTANDING SOCIODEMOGRAPHIC DISTRIBUTIONS OF HEALTH OVER TIME

David Rehkopf, Stanford University*

The standard assumption of linearity in the normal application of regression models is well known, despite the fact that a number of alternative models exist when this assumption is violated. There has been a lack of guidance on when these alternatives should be preferred. This talk will present a number of real scenarios where linear models provide a biased picture of the modeled associations, along with guidance on their evaluation. I will show how this bias can be evaluated and will present semi-parametric generalized additive mixed models to account for the potential non-linear dependence of outcome on exposure. The application comes from multiple waves of cross-sectional data spanning from the 1970s to the present from the National Health and Nutrition Examination Surveys. The analyses show that the shape of some associations of income and education with risk factors for cardiovascular disease vary dramatically over time. When examined using linear models, many of the critical associations are not apparent. By using multiple model fit statistics in combination with graphical representation of non-linearity over time, applied researchers will be better able to determine when a general class of models is more suitable.

e-mail: drehkopf@stanford.edu

ADJUSTING FOR MATCHING AND COVARIATES IN LINEAR DISCRIMINANT ANALYSIS

Josephine K. Asafu-Adjei, Harvard School of Public Health*
Allan R. Sampson, University of Pittsburgh
Robert A. Sweet, University of Pittsburgh

In studies that compare several diagnostic or treatment groups, subjects may not only be measured on a certain set of feature variables, but also matched on a number of demographic characteristics and measured on additional covariates. Data from multiple studies done on the same groups of subjects can be integrated and analyzed using statistical discrimination techniques, such as linear discriminant analysis, in order to identify which feature variables best discriminate among groups, while accounting for the dependencies among the feature variables. Subject matching and the use of covariates appear not to have been taken into consideration when implementing these discrimination methods (e.g., Knable et al., 2001). We present a modified approach to linear discriminant analysis that accounts for both covariate effects and the subject matching used in a particular study design. The methodology we develop is then applied to a series of post-mortem tissue studies conducted by Sweet et al. (2003, 2004, 2007, 2008) with the aim of comparing the neurobiological characteristics of subjects with schizophrenia to those of normal controls, and to a post-mortem tissue primate study conducted by Konopaske et al. (2008) comparing brain biomarker measurements across three treatment groups.

e-mail: jka7@pitt.edu

ADJUSTMENT FOR MEASUREMENT ERROR IN EVALUATING DIAGNOSTIC BIOMARKERS BY USING AN INTERNAL RELIABILITY SAMPLE

Matthew T. White, University of Pennsylvania*
Sharon X. Xie, University of Pennsylvania

Biomarkers are often measured with error due to imperfect lab conditions or temporal variability in subjects. Using an internal reliability sample of the biomarker, we propose a bias-correction approach for estimating a variety of diagnostic performance measures including sensitivity, specificity, the Youden index, the receiver operating characteristic curve, positive and negative predictive values, and positive and negative diagnostic likelihood ratios when the biomarker is subject to measurement error. We derive the asymptotic properties of the proposed likelihood-based estimators and show that they are consistent and asymptotically normally distributed. We demonstrate through simulations that the proposed approach removes the bias due to measurement error and outperforms the naive approach in estimating these diagnostic measures. We also derive the asymptotic bias of naive estimates. The proposed method has broad biomedical applications and is illustrated using a biomarker study in Alzheimer's disease. Using these data, we show that naive estimates of the diagnostic measures are biased towards estimates produced when the biomarker is ineffective. We recommend collecting an internal reliability sample during the biomarker discovery phase in order to adequately evaluate the performance of biomarkers with careful adjustment for measurement error.

e-mail: mwhi@mail.med.upenn.edu

INTEGRATING MULTIPLE MODALITIES OF HIGH THROUGHPUT ASSAYS USING ITEM RESPONSE THEORY: AN APPLICATION TO IDENTIFY GENES ALTERED IN OVARIAN CANCER

Pan Tong, University of Texas Health Science Center at Houston*
Kevin R. Coombes, University of Texas MD Anderson Cancer Center

Cancer is a complex disease that requires successive genetic and epigenetic alterations to achieve the hallmarks of cancer. Various mechanisms exist that can lead to gene dysfunction, including point mutation, copy number change, methylation, abnormal expression and so on. As a result, it makes more sense to identify altered genes by integrating different modalities of assays. In this project, we introduce the Item Response Theory (IRT) into bioinformatics research with a focus on identifying significantly altered genes in ovarian cancer patients. IRT, also known as latent trait theory or modern mental test theory, is a powerful method developed in psychometrics to construct, score and compare psychological and educational tests. By applying this method to a new setting and dealing with the computational challenge posed by high dimensional data, we are able to estimate the latent trait of alteration for individual genes from copy number, methylation and gene expression array data. It was found that severely altered genes supported by individual platforms differ. Further, novel significantly altered genes can be identified by integrating across platforms. The new method was also compared to conventional methods such as student's t-test and Wilcoxon rank-sum test. The result shows our method is more reliable and biologically meaningful than conventional methods.

e-mail: ptong1@mdanderson.org

ESTIMATING THE CORRELATION BETWEEN TWO VARIABLES SUBJECT TO LIMIT OF DETECTION

Courtney E. McCracken, Emory University
Stephen W. Looney, Georgia Health Sciences University*

Researchers are often interested in the relationship between biological concentrations obtained using two different assays, both of which may be biomarkers. Despite the continuing advances in biotechnology, the value of a particular biomarker may fall below some known limit of detection (LOD). Data values such as these are referred to as non-detects (NDs) and can be treated as left-censored observations. When attempting to measure the association between two concentrations, both of which are subject to NDs, serious complications can arise in the data analysis. Simple substitution, random imputation and maximum likelihood estimation methods are just a few of the methods that have been proposed for handling NDs when estimating the correlation between two variables, both of which are subject to left-censoring. Unfortunately, many of the popular methods require that the data follow a bivariate normal distribution or that only a small percentage of the data for each variable are below the LOD. These assumptions are often violated with biomarker data. In this presentation, we evaluate the performance of several methods, including Spearman's rho, when the data do not follow a bivariate normal distribution and when there are moderate to large censoring proportions in one or both of the variables.

e-mail: slooney@georgiahealth.edu

MODELING COMPLEX STRUCTURES IN NEUROPSYCHIATRIC TESTING DATA FOR SUBJECTS WITH PEDIATRIC DISORDERS

Vivian H. Shih, University of California at Los Angeles*
Laurie A. Brenner, University of California at Los Angeles
Carrie E. Bearden, University of California at Los Angeles
Catherine A. Sugar, University of California at Los Angeles
Steve S. Lee, University of California at Los Angeles

Recent debate in classical taxonomic categorizations of neuropsychiatric illnesses has sparked the field of phenomics – the study of dimensional patterns of deficits characterizing specific disorders. This area provides a rich opportunity for developing sophisticated statistical models due to multi-layer relationships between genes and behaviors, and complex interactions among neurocognitive measures. This talk focuses on temporal processing, a phenotype which may play a key role in language development and social communication. Approximately 300 children with ADHD, autism spectrum disorder (ASD), or 22q11.2 deletion syndrome completed a time reproduction task on five differently timed trials, repeated four times each in a random order. We explore temporal processing patterns in repeated measures mixed effects models, incorporating the effects of psychopathological symptoms such as inattention and hyperactivity on time reproduction. We further analyze the impact of temporal processing on social functioning. Since 22q11.2 deletion syndrome often coexists with ADHD and ASD, network models can uncover common phenotypes (ie: temporal processing, attentiveness, social reciprocity) that may be overlooked in traditional diagnostic definitions. The spectrum of symptom severity also creates complicated mixture distributions in the underlying data structure.

e-mail: vivianhshih@gmail.com

103. DYNAMIC TREATMENT REGIMENS

Q-LEARNING FOR ESTIMATING OPTIMAL DYNAMIC TREATMENT RULES FROM OBSERVATIONAL DATA

Erica E. Moodie, McGill University*
Bibhas Chakraborty, Columbia University

The area of dynamic treatment regimes (DTR) aims to make inference about adaptive, multistage decision-making in clinical practice. A DTR is a set of decision rules, one per interval of treatment where each decision is a function of treatment and covariate history which returns a recommended treatment. Q-learning is a popular method from the reinforcement learning literature that has recently been applied to estimate DTRs. While, in principle, Q-learning can be used for both randomized and observational data, the focus in the literature thus far has been exclusively on the randomized treatment setting. We extend the method to incorporate measured confounding covariates, using adjustment, propensity scores and inverse probability weighting. We summarize results of an extensive simulation study to compare different approaches to account for confounding in the Q-learning framework; the methods are examined under a variety of settings including practical violations of positivity and in nonregular scenarios. We illustrate the methods in examining the effect of breastfeeding on IQ in the PROBIT data.

e-mail: erica.moodie@mcgill.ca

WEIGHTED LOG-RANK STATISTIC TO COMPARE SHARED-PATH ADAPTIVE TREATMENT STRATEGIES

Kelley M. Kidwell, University of Pittsburgh*
Abdus S. Wahed, University of Pittsburgh

Adaptive treatment strategies more closely mimic the reality of a physician's prescription process where the physician prescribes a medication to his/her patient and based on that patient's response to the medication, modifies the treatment. Two-stage randomization designs, more generally, sequential multiple assignment randomization trial (SMART) designs, are useful to assess adaptive treatment strategies where the interest is in comparing the entire sequence of treatments, including the patient's intermediate response. In this paper, we introduce the notion of shared-path and separate-path adaptive treatment strategies and propose weighted log-rank statistics to compare overall survival distributions of two or more two-stage, shared-path adaptive treatment strategies. Large sample properties of the statistics are derived and the type I error rate and power of the tests are compared to standard statistics through simulation.

e-mail: kmk99@pitt.edu

A COMPARISON OF Q- AND A-REINFORCEMENT LEARNING METHODS FOR ESTIMATING OPTIMAL TREATMENT REGIMES

Phillip J. Schulte, North Carolina State University*
Marie Davidian, North Carolina State University
Anastasios A. Tsiatis, North Carolina State University

In clinical practice, physicians must make a sequence of treatment decisions throughout the course of a patient's disease based on evolving patient characteristics. At key decision points, there may be several treatment options and no consensus regarding which option is best. An algorithm for sequential treatment assignment at key decision points, based on evolving patient characteristics, is called a treatment regime. The statistical problem is to estimate the optimal regime which maximizes expected outcome. Q- and A-reinforcement learning are two methods that have been proposed for estimating the optimal treatment regime. While both methods involve developing statistical models for patient outcomes, A-learning is more robust, relaxing some assumptions. However, this additional robustness comes at a cost of increased variability and a bias-variance tradeoff between Q- and A-learning. We explore this tradeoff through parameter estimation and expected outcome for the estimated optimal treatment regime under various scenarios and degrees of model misspecification. We consider a setting of multiple treatment decisions over time with available observational data.

e-mail: pjschult@ncsu.edu

ESTIMATING INDIVIDUALIZED TREATMENT RULES USING OUTCOME WEIGHTED LEARNING

Yingqi Zhao, University of North Carolina at Chapel Hill*
Donglin Zeng, University of North Carolina at Chapel Hill
A. John Rush, University of North Carolina at Chapel Hill
Michael R. Kosorok, University of North Carolina at Chapel Hill

There is increasing interest in discovering individualized treatment rules for patients who have heterogeneous responses to treatment. In particular, one aims to find an optimal individualized treatment rule, which is a deterministic function of patient specific characteristics maximizing expected clinical outcome. In this paper, we first show that estimating such an optimal treatment rule is equivalent to a classification problem where each subject is weighted proportional to his or her clinical outcome. We then propose an outcome weighted learning approach based on the support vector machine framework. We show that the resulting estimator of the treatment rule is consistent. We further obtain a finite sample bound for the difference between the expected outcome using the estimated individualized treatment rule and that of the optimal treatment rule. The performance of the proposed approach is demonstrated via simulation studies and an analysis of chronic depression data.

e-mail: yqzhao@live.unc.edu

CHOICE OF OPTIMAL ESTIMATORS IN STRUCTURAL NESTED MEAN MODELS WITH APPLICATION TO INITIATING HAART IN HIV POSITIVE PATIENTS AFTER VARYING DURATION OF INFECTION

Judith J. Lok, Harvard School of Public Health*
Victor DeGruttola, Harvard School of Public Health
Ray Griner, Harvard School of Public Health
James M. Robins, Harvard School of Public Health

We estimate how a treatment effect depends on the time from infection to initiation of treatment, using observational data. A major challenge in making inferences from observational data is that treatment is not randomly assigned, which may induce bias in the effect estimate. We have developed a new class of Structural Nested Mean Models to estimate this effect. This leads to a large class of consistent, asymptotically normal estimators, under the assumption that all confounders are measured. However, estimates and standard errors turn out to depend significantly on the choice of estimators within this class, advocating the study of optimal ones. We will present an explicit solution for the choice of optimal estimators under some extra conditions. In the absence of those extra conditions, the resulting estimator is still consistent and asymptotically normal, although possibly not optimal. This estimator is also doubly robust: it is consistent and asymptotically normal not only if the model for treatment initiation is correct, but also if a certain outcome-regression model is correct. We illustrate our methods to investigate how the effect of initiating HAART in HIV-positive patients depends on the time between infection and treatment initiation in the early stages of infection.

e-mail: jlok@hsph.harvard.edu

104. MISSING DATA II

A JOINT LONGITUDINAL-SURVIVAL MODEL TO ANALYZE RISK FACTORS FOR DEATH OF PATIENTS ON THE LIVER TRANSPLANT WAITING LIST

Arwin Thomasson, University of Pennsylvania*
Peter Reese, University of Pennsylvania
David Goldberg, University of Pennsylvania
Sarah Ratcliffe, University of Pennsylvania

Data from waitlisted liver transplant patients is often analyzed by ignoring key features of this special type of patient. Longitudinal studies often overlook the non-random drop-out processes. Patients who are sicker may be more likely to drop out of the study due to health complications. Additionally, a level of health stability is needed for the patient to receive an eligible transplant and be censored out of the study. We present a novel joint model for biomarker trajectories and survival outcomes that takes into account both non-random drop-out processes. Both the biomarker trajectories and the survival process are modeled as non-linear functions with shared parameters. We demonstrate the advantages of this approach as compared with independent longitudinal and survival models. Our methods are applied to data from patients waitlisted for liver transplants at the Hospital of the University of Pennsylvania.

e-mail: arwin@mail.med.upenn.edu

MISSING COVARIATES AND THE PLAUSIBILITY OF THE MISSING AT RANDOM ASSUMPTION

Jonathan W. Bartlett, London School of Hygiene & Tropical Medicine, UK*
James R. Carpenter, London School of Hygiene & Tropical Medicine, UK
Kate Tilling, University of Bristol, UK
Michael G. Kenward, London School of Hygiene & Tropical Medicine, UK
Stijn Vansteelandt, Ghent University, Belgium

Missing data in covariates of a regression model is a common problem in epidemiological and clinical studies. A popular approach for handling such missingness is to use multiple imputation (MI), which assumes data are missing at random (MAR). We first argue that, in many settings, the MAR assumption is implausible for missing covariates, and that a more plausible assumption is that missingness is driven by the covariates themselves (and given these, not the outcome of interest). We examine the bias in an MAR analysis when in truth missingness depends only on the covariates, both analytically and through simulation. Under such an assumption for missingness, complete case analysis (CC) is valid, but potentially inefficient. We therefore propose estimators which aim to improve upon the efficiency of CC, and which involve specifying a parametric model for the missingness mechanism. We evaluate our proposed estimators in simulations and through application to an illustrative example.

e-mail: jonathan.bartlett@lshtm.ac.uk

MISSING VALUE IMPUTATION IN PHENOME DATA

Ge Liao, University of Pittsburgh*
George C. Tseng, University of Pittsburgh

In modern medical research, information of many clinical variables is routinely collected for each patient and analyzed. In the collection process of clinical data, missing values are inevitable. Such missing values can void application of subsequent statistical analysis since many methods require full data matrix in implementation. They may also distort the final conclusion of the analysis. Imputation of the missing values provides a reasonable solution in this situation. In this work, we will extend the K-nearest-neighbor (KNN) algorithm commonly used in microarray missing value im-

putation to impute missing values in phenome data. Distance measures between different types of variables are applied to identify nearest neighbors and missing values are estimated by regression methods. We develop KNN imputation by borrowing information from nearest subjects (KNN-S), from nearest variables (KNN-V) or their hybrid method (KNN-H). Their performances are evaluated by a standardized mean square error using simulated data and three large-scale phenome data sets in lung diseases and asthma.

e-mail: liaoge.serena@gmail.com

WEIGHTED SEMIPARAMETRIC ESTIMATION OF THE COX MODEL FOR INTERVAL-CENSORED DATA WITH MISSING COVARIATES

Lu Wang, University of Michigan*
Bin Nan, University of Michigan
Peng Zhang, Peking University
Andrew Zhou, University of Washington

Interval-censored time to event with missing covariates often arise in survey studies and many biomedical researches, especially when the disease is chronic and has long latent period before clinical symptoms. We consider a Cox's proportional hazard model and weight the log likelihood by the inverse selection probability (the probability of being observed) to adjust for missing covariates. We propose a spline-based sieve estimation approach, and maximize the weighted log likelihood in the sieve space where we approximate the log baseline hazard using B-spline functions. We show that both the estimates of the baseline hazard function and the regression parameters are consistent when the selection model is correctly specified. Asymptotic properties of the proposed estimator are developed. We derive the convergence rate for the sieve estimator of baseline hazard function, and establish root-n consistency of the regression coefficients. The finite sample performance of our proposed method is evaluated by simulation studies and the method is illustrated using data on development of dementia from the National Alzheimer's Coordinating Center

e-mail: luwang@umich.edu

GOODNESS-OF-FIT TEST TO DISTRIBUTION-FREE MODELS FOR LONGITUDINAL STUDIES WITH INFORMATIVE MISSING DATA

Pan Wu, University of Rochester*
Xin M. Tu, University of Rochester

Distribution-free models for longitudinal data have been used in a wide range of behavioral, psychotherapy, pharmaceutical drug safety, and health-care-related research studies. Existing theories for assessing the adequacy of model fit for parametric models and likelihood-based methods are no longer appropriate for their distribution-free counterparts. In addition, informative dropouts or study discontinuation are quite common in clinical trials and health-care studies. Ignoring missing values resulting from such premature study termination could cause biased estimation and reduce the power of hypothesis testing. A new approach of goodness-of-fit test is proposed to address missing data for distribution-free models. The proposed score-like test statistics are easy to apply, with nice asymptotic properties. The approach is illustrated with multiple simulation studies to demonstrate its superior performance over existing alternatives as assessed by both type I and type II error rates.

e-mail: pan_wu@urmc.rochester.edu

JOINT EMPIRICAL LIKELIHOOD CONFIDENCE REGIONS FOR THE EVALUATION OF CONTINUOUS-SCALE DIAGNOSTIC TESTS IN THE PRESENCE OF VERIFICATION BIAS

Binhuan Wang, Georgia State University*
Gengsheng Qin, Georgia State University

In a continuous-scale diagnostic test, the performance of the test to separate diseased subjects from non-diseased subjects can be measured by its specificity and sensitivity, when the cut-off level is given. Construction of joint confidence regions for specificity, sensitivity and a cut-off level could provide a solution to the selection a reasonable cut-off level. In some studies, not all subjects given their screening test results ultimately have their true disease status verified, and the verification may depend on the test result and the subject's observed characteristics. Directly applying current methods to verified subjects would result in verification-biased estimates. In this paper, by applying empirical likelihood method, a general framework combining empirical likelihood and general estimation equations with nuisance parameters is provided, and the asymptotic distribution of the empirical likelihood ratio statistics is obtained. This framework can be used to construct joint empirical likelihood confidence regions with verification-biased data. Simulation studies are conducted to evaluate the finite sample performance of the proposed confidence regions, and finally, a real example is given to illustrate the application of the new method.

e-mail: wang.binhuan@gmail.com

105. MULTIPLE TESTING

STEP-UP-DOWN MULTIPLE TESTING PROCEDURES AND THEIR CONTROL OF FALSE REJECTIONS

Alexander Y. Gordon, University of North Carolina at Charlotte*

Both classes of traditional step-down (Holm type) and step-up (Benjamini-Hochberg type) multiple testing procedures are contained in the class of step-up-down procedures introduced by Tamhane, Liu and Dunnett in 1998. The talk will focus on the exact levels at which any given step-up-down procedure controls the generalized family-wise error rates and the expected number of false rejections under a general and unknown dependence structure of individual tests. Explicit formulas for those levels will be given.

e-mail: aygordon@uncc.edu

AN IMPROVED HOCHBERG PROCEDURE FOR MULTIPLE TESTS OF SIGNIFICANCE

Dror M. Rom, PSI Center for Statistical Research*

A simple modification of Hochberg's (1988) step-up Bonferroni procedure for multiple tests of significance is proposed. The procedure is always more powerful than Hochberg's procedure, more powerful than Hommel's (1988) procedure for 3 and 4 tests, and strongly control the Family-Wise Error Rate. An extension to logically related hypotheses is demonstrated.

e-mail: d.rom@prosoftsoftware.com

ROBUST IDENTIFICATION OF CONDITIONAL GENE EXPRESSION IN DEVELOPMENT OF ONTHOPHAGUS BEETLES

Guilherme V. Rocha, Indiana University*
Karen Kafadar, Indiana University
Armin Moczek, Indiana University
Emilie Snell-Rood, University of Minnesota
Teiya Kijimoto, Indiana University
Justen Andrews, Indiana University

Multi-cellular organisms develop different tissues through cellular differentiation regulated by gene regulatory networks. Onthophagus taurus beetles stand out as a model organism in evolutionary developmental biology, due to the varied responsiveness of their phenotype to environmental factors, including the expression of horns: a novel complex trait with no homologous structure in other organisms. Identifying the genes involved in the differentiation of tissues according to gender and nutrition factors provides understanding of the molecular mechanisms involved in tissue development and offers insight into how novel traits might originate. A large microarray experiment was designed to assess the expression of genes in four tissue types of male and female beetles exposed to high and low levels of nutrition. We describe the analysis of the data from this study, which involves problems of multiple testing and estimating the relative sizes of differentially expressed genes under different conditions.

e-mail: gvrocha@indiana.edu

ESTIMATING THE NUMBER OF GENES THAT ARE DIFFERENTIALLY EXPRESSED IN BOTH OF TWO INDEPENDENT EXPERIMENTS

Megan C. Orr, Iowa State University*
Peng Liu, Iowa State University
Dan Nettleton, Iowa State University

A common procedure for estimating the number of genes that are differentially expressed (DE) in two experiments involves two steps. In the first step, data from the two experiments are separately analyzed to produce a list of genes declared to be DE in each experiment. Usually, each list is produced using a method that attempts to control the false discovery rate (FDR) in each experiment at some desired level alpha. In the second step, the number of genes common to both lists is counted and used as an estimate of the number of genes DE in both experiments. A problem with this approach is that the resulting estimates can vary greatly with alpha, and the values of alpha that produces the best estimate for any given pair of experiments is difficult to predict. We propose a method that uses the p-values from both experiments simultaneously to produce one estimate, which does not depend on FDR control, for the number of genes that are DE in both experiments. We compare the performance of our proposed method to that of the commonly used method with two simulation studies, one involving independent, normally distributed data and one involving microarray data. The results of the simulation studies demonstrate the advantages of our approach. We conclude the article by estimating the number of genes that are DE in both of two experiments involving gene expressions in maize leaves.

e-mail: meganorr@iastate.edu

AN ADAPTIVE RESAMPLING TEST FOR DETECTING THE PRESENCE OF SIGNIFICANT PREDICTORS

Ian W. 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 nonregular 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 nonregularity, 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 thresholding 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.

e-mail: mq2158@columbia.edu

JOINT MODELING OF MULTIPLE PARTIALLY OBSERVED OUTCOMES FROM CLINICAL TRIALS

Nicholas J. Horton, Smith College*
Kypros Kypri, University of Newcastle, Australia
Frank B. Yoon, Mathematica Policy Research
Garrett M. Fitzmaurice, Harvard Medical School
Stuart R. Lipsitz, Harvard Medical School
Sharon-Lise T. Normand, Harvard Medical School and Harvard School of Public Health

Prior work has shown that multiple outcomes are prevalent in many randomized trials, and that appropriate statistical analyses are often not used to account for this multiplicity. Interpretation of the results of trials with multiple outcomes is not always straightforward, particularly if some of the outcomes are sometimes missing. Joint tests, while not commonly used, are attractive in this setting because they capitalize on the correlation of multiple outcomes, provide a single and interpretable estimate of treatment effects, and can incorporate partially observed outcomes under certain assumptions regarding missingness. In this talk, we consider and implement principled methods for analyzing partially observed multiple outcome data arising from clinical trials using likelihood-based joint models. We use as a motivating example a trial which implemented a web-based intervention for n=7,237 university students in Australia, where 16% did not complete at least 1 follow-up assessment (unit non-response) and an additional number of outcomes were missing for some subjects (item non-response). Use of a joint model allowed incorporation of all observed information and yielded a clinically interpretable overall measure of treatment effectiveness.

e-mail: nhorton@smith.edu

A TIGHT PREDICTION INTERVAL FOR FALSE DISCOVERY PROPORTION UNDER DEPENDENCE

Shulian Shang, New York University*
Mengling Liu, New York University
Yongzhao Shao, New York University

Controlling the false discovery rate (FDR) has many important applications in genome-wide studies and other contexts where a large number of hypotheses are being tested simultaneously. In addition to controlling FDR, it is often desired to have an accurate prediction interval for the false discovery proportion (FDP). When common FDR control procedures are used, no tight prediction intervals for the FDP are currently available. We derive an explicit formula for the variance of FDP under general dependency among test statistics and obtain an upper prediction interval for the FDP under some semi-parametric dependence assumptions. Simulation studies indicate that the prediction intervals are tight with good coverage probabilities under weak dependence and for moderate sample size. We also present a permutation-based upper prediction interval for FDP which is useful when dependence is very strong. We illustrate the proposed prediction intervals using a publicly available prostate cancer dataset.

e-mail: ss4577@nyu.edu

106. POWER / SAMPLE SIZE CALCULATIONS

SAMPLE SIZE ESTIMATION IN RANDOMIZED CLINICAL TRIALS (RCTS) DESIGNED TO ESTABLISH THE INTERACTION BETWEEN PROGNOSTIC FACTOR AND TREATMENT: IMPACT OF PROGNOSTIC FACTOR DISTRIBUTION MISSPECIFICATION

William M. Reichmann, Boston University School of Public Health and Brigham and Women's Hospital*
Michael P. LaValley, Boston University School of Public Health
David R. Gagnon, Boston University School of Public Health
Elena Losina, Brigham and Women's Hospital and Boston University School of Public Health

Hypothesized interaction between treatment and prognostic factor in RCTs presents challenges for design and appropriate sample size estimation. Our objective was to examine the performance of sample size re-estimation (SSR) method for ensuring appropriate power and type I error under misspecification of the distribution of the prognostic factor. We examined the impact of the distribution of the dichotomous prognostic factor on power and sample size for the interaction effect. We varied the interaction magnitude, the prognostic factor distribution, and the magnitude and direction of the misspecification of the prognostic factor distribution. We examined SSR performance by conducting a simulation study. We compared empirical power and type I error, under different simulation scenarios of the pre-specified power (0.80) and type I error (0.05). Under no misspecification of the prognostic factor distribution, the SSR-based empirical power was greater than 80%. Negative misspecifications of the prognostic factor distribution decreased the power, estimated without SSR. SSR increased power but not always enough to reach the 80% target. SSR maintained a type I error rate below 5% for any misspecification of prognostic factor distribution under consideration. SSR can improve the power when the distribution of the prognostic factor is misspecified.

e-mail: breich@bu.edu

COMPARISON OF FOUR-PERIOD AND TWO-PERIOD CROSSOVER STUDIES FOR COMPARING WITHIN-SUBJECT VARIANCES OF TWO TREATMENTS

Donald J. Schuirmann, U.S. Food and Drug Administration*

In a comparative study such as a bioequivalence study, one possible reason for using a four-period crossover design, in which each subject receives a Test product (T) and a Reference product (R) twice, is to compare the within-subject variances of the two products. However, it has been implied (e.g. Gould 2000) that such a comparison may be made using a standard two-period crossover design, in which each subject receives each product only once. Guilbaud (1993) proposed methods for comparing the within-subject variances in a two-period crossover, under the assumption that the within-subject distribution, but not necessarily the between-subject distribution, is normal. In this presentation the efficiencies, for testing hypotheses about the ratio of within-subject variances, of the four-period and the two-period crossover designs are compared, both under the model assumed by Guilbaud and under a more general model described in the CDER guidance document Statistical Approaches to Establishing Bioequivalence (January 2001).

e-mail: donald.schuirmann@fda.hhs.gov

USE OF LONGITUDINAL REGISTRY DATA FOR OPTIMAL DESIGN OF CLINICAL TRIALS: AN EXAMPLE IN HUNTINGTON'S DISEASE

Elizabeth L. Turner, London School of Hygiene and Tropical Medicine*
Chris Frost, London School of Hygiene and Tropical Medicine

In the absence of data from randomised controlled trials it is informative to use observational data to inform clinical trial design. For a neurodegenerative disease such as Huntington's disease (HD) a successful treatment is one that alters the rate at which variables change over time. Outcome measures are therefore likely to be subject-specific slopes rather than absolute levels. Questions relating to the effect on power of extending trial follow-up, including multiple interim visits or using novel designs can be addressed by fitting linear mixed models (LMMs) to longitudinal data and the resultant estimates of between- and within-subject components of variance used to predict sample sizes for different designs (Frost, Kenward, Fox, Statist. Med.2008;27). The European HD Network Registry database includes longitudinal measures of many of the functional, cognitive, and motor score variables that are potential primary or secondary outcomes for clinical trials, with data for approximately 1500 people with HD seen typically annually for up to 5 years. We describe the LMMs used to model the longitudinal trajectories of Registry participants using the motor outcome as an example. Sample sizes were then estimated with bootstrapped confidence intervals to account for the uncertainty in the estimated variance parameters obtained from the LMMs.

e-mail: elizabeth.turner@lshtm.ac.uk

ASSESSING PROBABILITY OF SUCCESS FOR CLINICAL TRIALS WITH CORRELATED BINARY ENDPOINTS

Michael Dallas, Merck Research Laboratories*
Guanghan Liu, Merck Research Laboratories
Ivan Chan, Merck Research Laboratories
Joseph Heyse, Merck Research Laboratories

The probability of success (POS) is a measure of weighted power. The conventional statistical power for a clinical trial is calculated using a single assumed value for each parameter of interest and its variance, typically based on previous studies. To account for the uncertainty of these assumed values, POS is utilized, which incorporates prior distributions for the parameter estimate and/or its estimate of variance in a Bayesian framework. With multiple correlated endpoints, the calculation of POS becomes more complicated, especially in the case where there are many endpoints. Here, we focus on assessing POS for a clinical trial with multiple correlated binary endpoints. We examine several methods for calculating POS under this scenario, including a) using a specified prior distribution and b) using re-sampling to estimate the prior distribution. We apply the methods to a vaccine study with non-inferiority hypotheses based on multiple correlated binary endpoints.

e-mail: michael_dallas@merck.com

INTERIM DESIGN RESAMPLING FOR SAMPLE SIZE RE-ESTIMATION

Sergey Tarima, Medical College of Wisconsin*
Peng He, Medical College of Wisconsin
Tao Wang, Medical College of Wisconsin
Aniko Szabo, Medical College of Wisconsin

Internal pilot designs allow re-estimation of the sample size when an interim analysis is performed using available information. We re-sample the whole design at the interim analysis starting with a sample size recalculation based on the current observed values of the nuisance parameters and finishing with a decision to accept or reject the null. This internal re-sampling is performed under both the null and the alternative models to assess the bias of the type I error and power. We correct for this estimated on logit scale bias for interim sample size recalculation. We explore this suggested re-sampling approach under a set of simulation scenarios and compare its performance with several others previously published internal pilot designs.

e-mail: starima@mcw.edu

A GENERAL APPROACH FOR ESTIMATING STOPPING PROBABILITY OF LARGE CONFIRMATORY GROUP SEQUENTIAL CLINICAL TRIAL IN LIFE-THREATENING CONDITIONS MONITORING BINARY EFFICACY AND SAFETY OUTCOMES

Yanqiu Weng, Medical University of South Carolina*
Wenle Zhao, Medical University of South Carolina
Yuko Y. Palesch, Medical University of South Carolina

In large confirmatory trial in life-threatening conditions, some adverse event may not be rare or unexpected, and can be included in the formal sequential monitoring guideline. However, it is unknown to date how much impact safety monitoring would have on the error rates for efficacy analysis in this circumstance. On the other side, the decision making from the data safety monitoring committee (DMC) on early stopping a trial is very flexible, but current methods and available software for error spending and sample size calculation are all based on the assumption that DMC strictly follows the statistical guideline, which is unrealistic in practices. Based on these problems, we develop a new approach to estimate the marginal and joint stopping probabilities for efficacy and safety in large group sequential trial in life-threatening conditions. In addition to handling the multiplicity issue for multiple outcomes, the new approach is able to provide power estimations under various assumptions on data monitoring practices. The new approach is verified by Monte Carlo simulation and is demonstrated based on a real stroke trial. The result from this study suggests that formal safety monitoring in life-threatening conditions could have a dramatic impact on the error rates for efficacy analysis.

e-mail: weng@musc.edu

GEE METHOD FOR LONGITUDINAL DATA ANALYSIS IN SMART TRIALS AND THE ASSOCIATED SAMPLE SIZE FORMULA

Zhiguo Li, Duke University*

In sequnetial multiple assignment randomized trials (SMART), which are usually conducted in areas of chronic diseases or conditions, continuous logitudinal outcomes are frequently of primary interest. We consider the generalized estimating equation (GEE) approach to compare rates of change of repeated measurements under different adaptive treatment strategies. Inverse probability weighting is introduced to account for the fact that different subjects may have different probabilities of being consistent with a strategy. The asymptotic properties of the weighted GEE estimator of the parameters are obtained. And these properties are used to derive power and sample size formulae for this type of studies. The sample size calculation also takes into account the impact of missing data on the power. Simulation is conducted to assess the performance of the sample size formula in practical settings.

e-mail: zhiguo.li@duke.edu

107. IMAGING, OMICS, AND HIGH-DIMENSIONALITY

MULTIPLE COMPARISON PROCEDURES FOR iQTL ANALYSIS

Debashis Ghosh, Penn State University*
Wen-Yu Hua, Penn State University
Thomas E. Nichols, University of Warwick

Motivated by the the analysis of gene expression as a response variable in the analysis of quantitative trait loci, we consider in this talk the use of structural imaging profiles as a response variable in what we term imaging quantitative trait (iQTL) analyses. Thus, we are considering high-dimensionality of data on both the imaging and the genetic sides. The current standard has to perform massively univariate analyses that are termed voxel genomewide association studies (vGWAS). We will consider dimensionality reduction procedures on both the genetic and imaging sides and suggest new multiple testing procedures for selecting interesting genomic and imaging regions that will accommodate dependence. This will be achieved through two devices: the first is the idea of the "empirical null hypothesis" of Efron, while the second is a novel wavelet denoising procedure. Data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study will be used to illustrate the proposed methods.

e-mail: ghoshd@psu.edu

TEST FOR SNP-SET EFFECTS WITH APPLICATIONS TO SEQUENCING ASSOCIATION STUDIES

Xihong Lin, Harvard School of Public Health*

Sequencing studies are increasingly being conducted to identify rare variants associated with complex traits. The limited power of classical single marker association analysis for rare variants poses a central challenge in such studies. We propose a class of tests for testing for SNP set effects for association between genetic variants (common and rare) in a region and a continuous or dichotomous trait, while easily adjusting for covariates. We illustrate that they work well for a wide range of scenarios. Through analysis of simulated data across a wide range of practical scenarios and triglyceride data from the Dallas Heart Study, we show that SKAT can substantially outperform several alternative rare-variant association tests. We also provide analytic power and sample size calculations to help design candidate gene, whole exome, and whole genome sequence association studies.

e-mail: xhlin10@gmail.com

ANALYZING JOINT AND INDIVIDUAL VARIATION IN MULTIPLE DATA SETS

Andrew B. Nobel, University of North Carolina at Chapel Hill*
Eric S. Lock, University of North Carolina at Chapel Hill
J. S. Marron, University of North Carolina at Chapel Hill

In this talk we describe a method (JIVE) for the analysis of multiple, potentially high dimensional, data sets derived from a common set of samples, a setting similar to that of data fusion. The JIVE procedure decomposes multi-block data into a sum of three terms: a low-rank approximation capturing the joint variation across datatypes, a low-rank approximation for structured variation individual to each datatype, and residual noise. The JIVE decomposition can be used to quantify the amount of joint variation between datatypes, visually explore joint and individual structure, and reduce data dimensionality. The JIVE procedure is an extension of Principal Component Analysis (PCA) and in settings of interest has clear advantages over popular two-block methods such as Canonical Correlation Analysis (CCA) and Partial Least Squares (PLS). We will describe an application of JIVE to data from The Cancer Genome Atlas (TCGA).

e-mail: nobel@email.unc.edu

WHAT IS IN THE NEWS: AUTOMATIC AND SPARSE SUMMARIZATION OF LARGE DOCUMENT CORPORA

Luke Miratrix, University of California at Berkeley
Jinzhu Jia, Peking University
Brian Gawalt, University of California at Berkeley
Laurent El Ghaoui, University of California at Berkeley
Jas Sekhon, University of California at Berkeley

In this talk, I will introduce the statnews project at UC Berkeley in collaboration with the El Ghaoui group in EE. Our goal is an automatic and interpretable summarization of large document corpora on a subject/topic (e.g., "China") to allow social scientists screen large bodies of text and to suggest further readings. We encode each text unit as its uni-gram, bi-gram and tri-gram counts (e.g., article or paragraph as a unit) to obtain large and sparse matrices. We adopt the predictive framework in machine learning by automatically labeling text units as positive or negative examples according to appearances of phrases related to the subject. After data preprocessing, we employ computationally feasible sparse feature selection methods to derive lists of words/phrases that associate with a particular subject matter (e.g. "China") in, for example, the New York Times International Section. We designed and carried out a human experiment to compare different efficient feature selection methods including Lasso and L1 penalized Logistic regression. Lasso is found to be a good overall method and L2 normalization seems well-suited for short units of text such as paragraphs. If time allows, I will present results from an on-going project using statnews tools to answer questions in political science.

e-mail: binyu@stat.berkeley.edu

108. STATISTICAL METHODS FOR MODELING SEER POPULATION-BASED CANCER DATA

INTRODUCTION: AN OVERVIEW OF POPULATION-BASED SEER CANCER REGISTRY DATA

Hyunsoon Cho, National Cancer Institute, National Institutes of Health*
Nadia Howlader, National Cancer Institute, National Institutes of Health

The Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute is a geographic area based cancer registry in the United States. SEER collects and publishes cancer incidences, prevalence and survival data annually, and SEER is a unique source of population-based cancer data used by researchers, clinicians, legislators, health planners, public health officials, patients etc. For over six million patients, SEER Research Database has demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status information. In this talk, we present an overview of SEER data. A summary of methods and associated software tools available for analysis and reporting of population-based cancer statistics such as incidence, mortality, survival, prevalence, and spatial statistics will also be discussed.

e-mail: hyunsoon.cho@nih.gov

USING SEER DATA TO DEVELOP MODELS OF ABSOLUTE CANCER RISK

Mitchell H. Gail, National Cancer Institute, National Institutes of Health*

Researchers who want to develop models of the absolute risk of cancer incidence or of the absolute risk of dying of a given cancer following cancer diagnosis can use data available in the National Cancer Institute's Surveillance, Epidemiology and End Results (SEER) program. If population-based case-control data for a specific cancer are available, this information can be combined with age- specific incidence data on that cancer from SEER and with age-specific mortality rates for competing causes of death to compute absolute risk. SEER follow-up information on persons di-

agnosed with cancer can be used directly to estimate the absolute risk of dying of that cancer, taking into account the chance that the person will die of some other illness first. The SEER data are also useful for studying the absolute risk of a second cancer following incidence of a first cancer. This talk will illustrate these applications of SEER data.

e-mail: gailm@mail.nih.gov

DETECTING MULTIPLE CHANGE POINTS IN PIECEWISE CONSTANT HAZARD FUNCTIONS

Yi Li, University of Michigan*
Mitchell H. Gail, National Cancer Institute, National Institutes of Health
Melody Goodman, Wasington University, St. Louis

The National Cancer Institute (NCI) suggests a sudden reduction in prostate cancer mortality rates, likely due to highly successful treatments and screening methods for early diagnosis. We are interested in understanding the impact of medical breakthroughs, treatments, or interventions, on the survival experience for a population. For this purpose, estimating the underlying hazard function, with possible time change points, would be of substantial interest, as it will provide a general picture of the survival trend and when this trend is disrupted. Increasing attention has been given to testing the assumption of a constant failure rate against a failure rate that changes at a single point in time. We expand the set of alternatives to allow for the consideration of multiple change-points, and propose a model selection algorithm using sequential testing for the piecewise constant hazard model. These methods are data driven and allow us to estimate not only the number of change points in the hazard function but where those changes occur. Such an analysis allows for better understanding of how changing medical practice affects the survival experience for a patient population. We test for change points in prostate cancer mortality rates using the NCI Surveillance, Epidemiology,and End Results dataset.

e-mail: yili@umich.edu

MAMMOGRAPHY, MODELING AND POLITICS

Jeanne Mandelblatt, Lombardi Cancer Center at Georgetown University*
Kathy Cronin, National Cancer Institute, National Institutes of Health
Don Berry, University of Texas MD Anderson Cancer Center
Harry DeKoning, Erasmus University Medical Center
Sandra Lee, Harvard University
Sylvia Plevritis, Stanford University
Clyde Schechter, Albert Einstein College of Medicine
Natasha Stout, Harvard Pilgrim Healthcare
Marvin Zelen, Harvard University
Eric Feuer, National Cancer Institute, National Institutes of Health

The optimal schedule for breast cancer screening remains controversial. Conducting new trials to identify precise mortality benefits for women of different age groups is not feasible. Therefore, the Cancer Intervention and Surveillance Modeling Network (CISNET) applied established models to estimate the outcomes expected from breast cancer screening. Briefly, 6 models were used to estimate the outcome from 20 mammography screening strategies used varied age of initiation and cessation or screening, and interval of screening. National data on age-specific incidence of cancer, mammography characteristics, treatment effects and competing mortality were used. The models assume 100% adherence to screening and treatment. The modeling groups concluded that the most efficient screening strategies are those that include a biennial screening interval. Decisions about optimal starting and stopping age depend on willingness to accept false-positive results and potential over diagnosis. The CISNET group made no judgments about screening policy, and stressed that decisions about the optimal screening strategy depend on programmatic goals and the objectives of the individual related to benefits and potential harms. This presentation will summarize the findings from this modeling project as it pertains to recommendations regarding screening mammography and highlight the political responses to the research.

e-mail: mandelbj@georgetown.edu

109. POWERFUL STATISTICAL MODELS AND METHODS IN NEXT GENERATION SEQUENCING

ANALYTICAL CHALLENGES IN ASSOCIATION STUDIES WITH WHOLE-GENOME SEQUENCING

Dan L. Nicolae, University of Chicago*

We have started the transition from single-marker tests on common SNPs in genome-wide association studies to set-based inference on all variants in a functional element, with genotypes called from second-generation sequencing data. We discuss here some of the challenges in this transition, including: (i) the construction of sets; (ii) the implicit and explicit assumptions on underlying genetic models of risk for a given set; (iii) interactions with environment and ancestry; and (iv) the interpretation of results.

e-mail: nicolae@galton.uchicago.edu

ASSOCIATION ANALYSIS OF GENOME-WIDE GENETIC DATA

Li Hsu, Fred Hutchinson Cancer Research Center*

Genome-wide association studies (GWAS) have successfully identified hundreds of novel variants that are associated with complex diseases; however, much of the inheritable disease variation is still unexplained. It has been suggested that the missing heritability may be due to rare genetic variants that GWAS marker panels do not cover, gene-environment interaction, gene-gene among others. For these, power is a critical issue, as generally more subjects are needed to detect rare variants association or gene-environment and gene-gene interactions. In this talk, I will present methods that would enhance power for these analyses. Simulation results and real data examples will be used to illustrate the methods.

e-mail: lih@fhcrc.org

EXPONENTIAL COMBINATION PROCEDURE FOR SET-BASED TESTS IN SEQUENCING STUDIES

Lin S. Chen, University of Chicago*
Li Hsu, Fred Hutchinson Cancer Research Center
Dan L. Nicolae, University of Chicago

Next-generation sequencing studies provide an in-depth exploration of human genome. To identify genetic factors that are associated with disease risk, methods have been proposed to jointly analyze variants in a set (e.g., a gene). Variants in a properly defined set could be associated with disease risk concertedly and by accumulating information among them, power to detect genetic risk factors may be improved. Most set-based methods in the literature have set statistics that can be written as the linear summation of variant-based statistics within the set. We propose the exponential combination of variant based statistics as a Bayes test for sparse alternative, to account for the distinctive features of sequencing data. That is, the risk-associated variants are relatively sparse compared with the large number of variants being tested in a set. We derive the exponential combination (EC) tests for three existing set-based methods and show by simulations and on the 1000 Genomes data set that EC greatly improves power.

e-mail: lchen11@uchicago.edu

110. RECENT ADVANCES IN CLINICAL TRIAL DESIGN: UTILITIES AND PITFALLS

COMPLEX CLINICAL TRIAL DESIGNS: AN OVERVIEW

H.M. James Hung, U.S. Food and Drug Administration*

In many regulatory applications, the designs of pivotal clinical trials are increasingly more complex because of many necessary considerations, such as cost, ethics and efficiency. In psychiatry trials where the placebo response is high, the trial design may need to incorporate some kind of enrichment strategy to reduce the number of placebo responders. In large clinical outcome cardiovascular trials, adaptive selection designs may need to be considered to study multiple doses or multiple patient groups in one trial. The selection may be based on immediate endpoints or markers which have not been proven as surrogate endpoints. When placebo cannot be used, an active control or historical control may have to be used and consequently statistical inferences would have to rely on indirect or nonrandomized comparisons. This presentation will give an overview of some of the complex trial designs in terms of utilities, pitfalls and challenging issues. Statistical inference frameworks will be discussed.

e-mail: hsienming.hung@fda.hhs.gov

A DOUBLY ENRICHED CLINICAL TRIAL DESIGN MERGING PLACEBO LEAD-IN AND RANDOMIZED WITHDRAWAL

Roy N. Tamura, Eli Lilly and Company*
Anastasia Ivanova, University of North Carolina at Chapel Hill

A new clinical trial design, designated the doubly enriched design (DED), is introduced which augments the standard randomized placebo controlled trial with second stage enrichment designs in placebo non-responders and drug responders. The trial is run in two stages. In the first stage patients are randomized between drug and placebo. In the second stage placebo non-responders are re-randomized between drug and placebo and drug responders are re-randomized between drug and placebo. All first stage data, and second stage data from first stage placebo non-responders and first stage drug responders, are utilized in the efficacy analysis. We illustrate the use of one, two, and three degrees of freedom score tests for the DED. An illustration of the DED is given for generalized anxiety disorder and compared to the standard parallel clinical trial, placebo lead-in and randomized withdrawal designs in terms of sample size and total patient exposure time.

e-mail: tamura_roy_n@lilly.com

UTILITY AND PITFALLS WITH ADAPTIVE SELECTION DESIGN

Sue-Jane Wang, U.S. Food and Drug Administration*

An adaptive design can be as simple as a preplanned increase in sample size to as complex as a preplanned modification of design aspects including sample size, treatment dose(s), patient subsets, study objective, study duration, etc. An adaptive design may combine stages of a trial or phases of clinical trials, making it a single controlled trial. There is a genuine interest in combining the early data that is used to learn with the independent data to confirm the hypothesis selected based on the interim learning data, though there are concerns of doing so. Proposals of adaptive design are increasing. This presentation will introduce adaptive selection designs including the possibility to increase the sample size under adaptive selection. The utility and pitfalls with use of adaptive selection in confirmatory setting will also be discussed in light of regulatory science.

e-mail: suejane.wang@fda.hhs.gov

A TWO PART DESIGN FOR EVALUATING ANTIEPILEPTIC DRUGS

Eugene Laska, New York University School of Medicine*

In many diseases placebo can cause so much harm that its use as a control in a RCT is not possible. Various approaches are used to get around the difficulty that this imposes on interpretation of clinical trial results, but its impact on drug development is considerable. It is generally agreed that it is unethical to use a placebo in evaluating antiepileptic drugs. AEDs are first evaluated as add-on, or adjunctive therapies, where test T or placebo P are randomly added to existing treatments. In the past, monotherapy trials used pseudo-placebos, low doses of active drugs. There are now 17 treatments approved for adjunctive use in partial-onset seizures. Of the 12 AEDs approved in the last 20 years, only four are indicated for monotherapy. The proposed two-part design tests both adjunctive and monotherapy. The first stage is the standard add-on trial where T and P are randomly added to (perhaps only one) background AED in the usual fashion. In the second enrichment stage, responders on the T arm randomly remain on their background AED or have it replaced by P. Those on T plus placebo would be appraised against those on T plus their background AED with a non-inferiority approach.

e-mail: laska@nki.rfmh.org

111. INDIVIDUALIZED RISK PREDICTION USING JOINT MODELS OF LONGITUDINAL AND SURVIVAL DATA

A JOINT MODEL OF CERVICAL CANCER, PAP SMEARS, AND HPV TESTS FOR USE IN DEVELOPING CANCER SCREENING GUIDELINES

Hormuzd A. Katki, National Cancer Institute, National Institutes of Health*
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Cancer screening guidelines should be based on the cancer risk implied by each combination of screening tests. However, risk is difficult to estimate from screening data where decisions to treat are strongly based on the history of screening test results and where participants may not comply with screening protocols. Using data on over 330,000 women undergoing cervical cancer screening, we modeled cervical cancer risk along with longitudinal models of Pap smears and human papillomavirus (HPV) acquisition/persistence/clearance (and relevant demographic variables). Particular statistical challenges we faced were modeling multivariate discrete longitudinal processes and informative censoring due to non-compliance with screening protocols. We suggest how risk estimates from this model might inform cervical cancer screening guidelines.

e-mail: katkih@mail.nih.gov

PREDICTION OF MULTIVARIATE BINARY DATA WITH MULTI-SCALE INFORMATIVE DROPOUT—A JOINT MODELING APPROACH

Alexander C. McLain, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health*
Rajeshwari Sundaram, Eunice Kennedy Shriver National Institute of Child Health and Human Development, National Institutes of Health

Prospective pregnancy studies provide a wealth of information regarding the behaviors of couples attempting to achieve pregnancy. One aspect of particular interest is using the prospective pregnancy studies to develop a model for individualized prediction of time-to-pregnancy (TTP). An integral part of creating accurate predictions of TTP is in predicting a couples' intercourse behavior. In prospective pregnancy studies, intercourse behavior is collected on each day in a woman's cycle, for as many cycles as it takes for a woman to get pregnant. The number of days where an intercourse can occur is governed by two processes: the variable length of the menstrual cycles and the number of cycles observed. We discuss how key concepts of missing at random and missing not at random translate when the missing data mechanism

is multi-scaled, and how these relate to prediction of the outcome. Specifically, we demonstrate through theory and simulation the effects of accounting for multiple missing data mechanisms on estimation and prediction. We present an analysis of the Stress and Time-to-Pregnancy subcomponent of the Oxford Conception study. Within this analysis we quantify the effect accounting for various missing data mechanisms has on prediction of the intercourse values in terms of the risk for subinfertility (a TTP> 6 cycles).

e-mail: mclaina@mail.nih.gov

DIFFERENT PARAMETERIZATIONS FOR JOINT MODELS FOR LONGITUDINAL AND SURVIVAL DATA, AND HOW THEY AFFECT INDIVIDUALIZED PREDICTIONS

Dimitris Rizopoulos, Erasmus University Medical Center*

Recently there has been an increasing interest in using joint models for longitudinal and event time data to obtain individualized predictions of survival probabilities. The intriguing feature of this application of joint models is that survival probabilities are dynamically updated as extra longitudinal information is collected for the subject(s) of interest. In this work we explore different parameterizations for the association structure between the longitudinal and event time outcomes, and we investigate how these can affect predictions.

e-mail: d.rizopoulos@erasmusmc.nl

JOINT LATENT CLASS MODELS OF LONGITUDINAL AND TIME-TO-EVENT DATA IN THE CONTEXT OF INDIVIDUAL DYNAMIC PREDICTIONS

Cécile Proust-Lima, INSERM, France*
Mbéry Séne, INSERM, France
Jeremy MG Taylor, University of Michigan
Hélène Jacqmin-Gadda, INSERM, France

Most statistical developments in joint modelling area have focused on the shared random-effect models that include characteristics of the longitudinal marker as predictors in the model for the time-to-event. A less well known approach is the joint latent class model which consists in assuming that a latent class structure entirely captures the correlation between the longitudinal marker trajectory and the risk of event. Thanks to its flexibility in the modelling of the dependency between the longitudinal marker and the time-to-event, as well as in the covariates inclusion, the joint latent class model may be particularly suited for prediction problems. We aim at giving an overview of joint latent class modelling, especially in the prediction context. We introduce the model, its estimation and its evaluation of goodness-of-fit, as well as the main differences with the shared random-effect model. Then, dynamic predictive tools derived from joint latent class models are presented, as well as measures to evaluate their dynamic predictive accuracy. A detailed illustration of the methods is given in the context of the prediction of Prostate cancer recurrence after radiation therapy based on repeated measures of Prostate Specific Antigen.

e-mail: cecile.proust@isped.u-bordeaux2.fr

112. RECENT ADVANCES IN DYNAMIC TREATMENT REGIMES RESEARCH

PRACTICAL ISSUES IN THE DESIGN, CONDUCT, AND ANALYSIS OF RANDOMIZED ONCOLOGY TRIALS COMPARING DYNAMIC TREATMENT REGIMES

Peter Thall, University of Texas MD Anderson Cancer Center*

In oncology, a common dynamic treatment regime (DTR) is a pair (A,B), where A is an initial frontline treatment and B is a salvage treatment given if A fails due to disease worsening, severe toxicity, or failure to achieve a remission. This paradigm may be extended to include additional stages, since a patient's cancer may repeatedly respond to therapy with the disease later worsening, and it applies to many other disease areas. Evaluating effects of frontline treatments on survival while ignoring effects of salvage therapies, or conversely in phase II trials focused on salvage therapies, muddies the combined effect of the DTR. Consequently, results of conventional trials based on either frontline alone or salvage alone are of limited use to clinical oncologists because they ignore actual medical practice. Recognition of these facts has motivated trials at M.D. Anderson Cancer Center that aim to compare two-stage DTRs. Designing, organizing, conducting, and analyzing these trials has proved to be extremely challenging. In this talk, I will discuss my experiences with oncology trials that I have designed: a completed trial of 12 multi-stage DTRs for advanced prostate cancer, and an ongoing trial of 6 DTRs for metastatic renal cancer.

e-mail: rex@mdanderson.org

A POLICY SEARCH METHOD FOR ESTIMATING TREATMENT POLICIES

Xi Lu, University of Michigan*
Susan A. Murphy, University of Michigan

A treatment policy or dynamic treatment regime is a sequence of decision rules. At each stage a decision rule inputs patient history and outputs a treatment. The value of a treatment policy is the expected outcome when the policy is used to assign treatment. Data from sequential, multiple assignment, randomized trials can be used to estimate an optimal treatment policy. We propose a new method for estimating the expected outcome of a policy. In this new method each stage's treatment effect or "blip" is parameterized. These treatment effects are easily interpretable to scientists and thus meaningfully parameterized. To estimate the parameters we utilize a telescoping sum representation of the policy value and employ ideas from missing data theory. We illustrate the proposed method with data from the ExTENd trial, a recently completed alcohol dependence study.

e-mail: luxi@umich.edu

COMPARING DYNAMIC TREATMENT REGIMES VIA THE G-FORMULA

Miguel A. Hernan, Harvard School of Public Health*

The estimation of causal effects of dynamic treatment regimes requires the use of methods that appropriately adjust for time-varying confounding. Three classes of methods developed by Robins and collaborators can be used for this purpose: inverse probability weighting of dynamic marginal structural models, g-estimation of structural nested models, and the parametric g-formula. This talk describes recent advances in the practical implementation of the parametric g-formula (including software developments), and will discuss the relative advantages and disadvantages of the g-formula compared to inverse probability weighting and g-estimation.

e-mail: mhernan@hsph.harvard.edu

REALISTIC AS TREATED DYNAMIC TREATMENT REGIMES

Andrea Rotnitzky, Universidad Di Tella and Harvard University*
Sebastien Haneuse, Harvard School of Public Health

We consider a point exposure study in which treatment doses can take values in an interval (a,b) of the real line. Motivated by the problem of estimating optimal realistic reductions in surgical times of lung cancer patients that minimize the probability of short term post-surgical complications, we consider a dynamic treatment regime in which each subject receives a reduction of the surgical time (dose) that he/she actually received, where the reduction may depend on the subject's own covariates. The estimand of interest is the effect of the probability of the counterfactual post-surgical complication under the aforementioned dynamic treatment regime. We develop outcome regression, inverse probability weighted and double-robust estimators of the estimand. We argue that, in contrast to estimands of population average effects, our IPW estimator of a continuous treatment is stable. In addition, by virtue of the fact that our estimand is a functional of the treatment process, the double-robust estimator has lower asymptotic efficiency than the locally efficient inverse probability weighted estimator. This work is joint with Sebastien Haneuse.

e-mail: arotnitzky@utdt.edu

113. A REVIEW OF ESTABLISHED AND NEW METHODS OF MULTIPLE IMPUTATION OF MISSING DATA WITH THE EMPHASIS ON AVAILABLE SOFTWARE PACKAGES

FLEXIBLE IMPUTATION WITH MICE

Stef van Buuren, TNO*

The MICE algorithm imputes data through an imputation model that consists of a set of conditional distributions. This type of model allows for imputations that are close to the data. The methodology has been expanded and refined in my new book "Flexible Imputation of Missing Data", first available at ENAR 2012. The lecture reviews state-of-the-art tools and techniques to create and evaluate plausible multiple imputations, and shows practical application is possible using the mice package in R.

e-mail: stef.vanbuuren@tno.nl

MULTIPLE IMPUTATION BY ORDERED MONOTONE BLOCKS WITH APPLICATION TO THE ANTHRAX VACCINE ADSORBED TRIAL

Fan Li, Duke University*
Michela Baccini, University of Florence
Fabrizia Mealli, University of Florence
Constantine Frangakis, Johns Hopkins University
Elizabeth Zell, Centers for Disease Control and Prevention
Donald B. Rubin, Harvard University

Multiple imputation (MI) has become a standard statistical technique for imputing missing values, where imputations are created as random draws from the posterior predictive distribution of the missing data. The Anthrax Vaccine Adsorbed (AVA) trial data created new challenges for MI due to the large number of variables of different types and the limited sample size. An intuitive method for handling such complex data is to specify, for each variable with missing values, a univariate conditional distribution given all other variables, in the form of a regression model. Such univariate imputation strategies are valid for monotone missing data, but have the theoretical drawback that the fully conditional distributions are generally incompatible when missing data are not monotone. Aiming at reducing incompatibility, we propose the "multiple imputation by ordered monotone blocks" approach to extend the theory for monotone patterns to arbitrary missing patterns. The key idea is to break an arbitrary missing pattern into a collection of smaller but monotone missing patterns. We apply this strategy to impute the missing data in the AVA trial and evaluate its performance by a novel simulation-based approach. A method for creating missing values in the simulated data sets, which mimics the observed missing data patterns, is also proposed.

e-mail: fli@stat.duke.edu

NEW MULTIPLE IMPUTATION METHODS IN SOLAS, INCLUDING A COMBINATION OF TWO HOT-DECK METHODS WITH APPEALING PROPERTIES

Donald B. Rubin, Harvard University
Victoria Liublinska, Harvard University*

Increasing numbers of researchers are making efforts to address the issue of missing data in their work. The failure to do so may result in biased conclusions or at inefficient analyses. SOLAS software package offers a variety of statistical methods that address missingness. We will concentrate on two hot-deck methods of data imputation with different objectives. Propensity Score Matching method (PSM), introduced in R. Little (1986), is aimed at reducing bias when estimating the average of the (partially missing) variable of interest. Predictive Mean Matching (PMM), first proposed by D. Rubin (1986), captures the relationship between the variable of interest and other more observed variables. We study large-sample properties of both methods and ways to combine them to achieve trade-off between the two approaches, especially in a setting when one or both models are misspecified.

e-mail: vliublin@fas.harvard.edu

CONVERGENCE PROPERTIES OF SEQUENTIAL REGRESSION MULTIPLE IMPUTATION APPROACH

Trivellore Raghunathan, University of Michigan*
Jian Zhu, University of Michigan

A sequential regression approach (also called chained equations) is an attractive method for multiply imputing missing values in a complex data base with skip patterns, bounds and other restrictions. In this approach, imputations are carried out by using a Gibbs sampling type iterative algorithm based on specifying a series of conditional distributions for each variable conditional on all other variables. The imputations are the draws from the corresponding posterior predictive distribution of the missing values. Since the specification of just conditional distributions does not guarantee the existence of a joint distribution, the traditional convergence results for Gibbs sampling algorithm does not apply. In this paper, we will investigate the consequences of such incompatibility where the sequence of regression model fits the data well but a joint distribution does not exist. We show, through theoretical development and simulation studies, that multiple imputation using incompatible distributions provide valid completed-data inferences.

e-mail: teraghu@umich.edu

MAKING MULTIPLE IMPUTATION ACCESSIBLE TO NON-STATISTICIANS

Leland Wilkinson, University of Illinois at Chicago*

Twenty-five years after Rubin's influential book introducing multiple imputation (MI), we now have numerous computer programs implementing MI for a variety of statistical models. Unfortunately, most of these programs have two drawbacks: they are inaccessible to ordinary users and they are relatively limited in scope. They are inaccessible because they usually require generating imputations, saving results into files, and merging files to combine imputations. They are limited in scope because they usually cover only simple regression models. While Rubin's algorithm presents few computational difficulties, the real problem for software developers is that MI needs to be embedded at the heart of all statistical calculations. This requires a redesign of the architecture of the traditional statistics package, including more flexible statistical programming systems such as R and Stata. A new computational platform, called A Second Opinion, illustrates the steps necessary to place MI at the most fundamental level of statistical calculations.

e-mail: leland.wilkinson@gmail.com

114. ACCELERATED FAILURE TIME MODELS

ACCELERATED FAILURE TIME MODEL FOR CASE-COHORT DESIGN WITH LONGITUDINAL COVARIATES MEASURED WITH ERROR

Xinxin Dong, University of Pittsburgh*
Lan Kong, Penn State Hershey College of Medicine
Abdus S. Wahed, University of Pittsburgh

Repeated measurements of biomarkers are often assembled at informative observation times to better understand the mechanism of a disease. In large cohort studies, it is prohibitive to measure multiple candidate markers over time for every subject. Case-cohort design provides a cost effective solution when the markers of interest are expensive to measure and/or the event rate is low. Under a case-cohort design, biomarkers are only measured for a subcohort that is randomly selected from the entire cohort at the beginning of the study, and any additional cases outside the subcohort. To reveal the relationship between biomarker trajectories and the time to event from case-cohort studies, we propose a joint analysis approach to account for the longitudinal covariates measured with error in the accelerated failure time (AFT) model. The maximum likelihood estimators are obtained by Gaussian quadrature method. We evaluate the performance of our case-cohort estimator and compare its relative efficiency to the full cohort estimator through simulation studies. The proposed procedure is further demonstrated using the data from a biomarker study of sepsis among patients with community acquired pneumonia.

e-mail: eva.dongxinxin@gmail.com

ACCELERATED FAILURE TIME MODELING OF GENETIC PATHWAY DATA USING KERNEL MACHINES FOR RISK PREDICTION

Jennifer A. Sinnott, Harvard University*
Tianxi Cai, Harvard University

Integrating genomic information with traditional clinical risk factors to improve the prediction of disease outcomes could profoundly change the practice of medicine. However, the large number of potential markers and the complexity of the relationship between markers and disease make it difficult to construct accurate risk prediction models. Standard approaches to identifying important markers often rely on marginal associations and may not capture non-linear or interactive effects. At the same time, much work has been done to group genes into pathways and networks. Integrating this kind of prior biological knowledge into statistical learning could potentially improve model interpretability and reliability. One effective approach is to employ a kernel machine (KM) framework, which has been recently extended to analyzing survival outcomes under the Cox model. In this paper, we propose KM regression under an accelerated failure time model. We derive a pseudo score statistic for testing and a risk score for prediction survival. To approximate the null distribution of our test statistic, we propose resampling procedures which also enable us to develop alternative robust testing procedures that combine information across models and kernels. Numerical studies show that the testing and estimation procedures perform well. The methods are illustrated with an application in breast cancer.

e-mail: jsinnott@hsph.harvard.edu

A SEMIPARAMETRIC ACCELERATED FAILURE TIME PARTIAL LINEAR MODEL AND ITS APPLICATION TO BREAST CANCER

Yubo Zou, University of South Carolina
Jiajia Zhang, University of South Carolina*
Guoyou Qin, Fudan University, Shanghai, PR China

Breast cancer is the most commonnon-skin cancer in women and the second most common cause of cancer-related death in US women. It is well known that the breast cancer survivalrate varies with age at diagnosis. For most cancers, the relative survival rate decreases with age, but breast cancer may show an unusual age pattern. In order to reveal the stage risk and age effects pattern, we propose a semiparametric accelerated failure time partial linear model and develop its estimation method based on the penalized spline (P-spline) and the rank estimation approach. The simulation studies demonstrate that the proposed method is comparable to the parametric approach when data is not contaminated, and more stable than parametric methods when data is contaminated. By applying the proposed model and method to the breast cancer data set of Atlantic County, NewJersey, from the SEER program, we successfully reveal the significant effects of stage, and show that women diagnosed at age around 38 years have consistently higher survival rates than either younger or older women.

e-mail: jzhang@mailbox.sc.edu

PARAMETRIC INFERENCE ON ACCELERATED FAILURE TIME MODEL WITH RANDOM EFFECTS

KyungAh Im, University of Pittsburgh*
Jong-Hyeon Jeong, University of Pittsburgh
Rhonghui Xu, University of California-San Diego

We propose a parametric inference on an accelerated failure time (AFT) model with random effects for correlated or clustered survival data through the full likelihood function. We assume that the error distribution for the AFT model belongs to the G-rho family of Harrington and Fleming (1982, Biometrika 69, 553-566) and the random effects follow the multivariate normal distribution. A modified Expectation-Maximization (EM) algorithm will be used to maximize the observed likelihood in the presence of random effects. The conditional expectation in E-step was computed by Markov Chain Monte Carlo (MCMC) method using Adaptive Rejection Metropolis Sampling (ARMS) within Gibbs sampling (Gilks et al., 1995, Applied Statistics, 44, 455-472). Simulation studies are performed to assess the estimates of the fixed and random effects in the model for various scenarios of cluster size and number of subjects in each cluster. The proposed method is applied to a real dataset from a breast cancer clinical trial.

e-mail: kellyim1@gmail.com

BAYESIAN SEMIPARAMETRIC ACCELERATED FAILURE TIME MODEL FOR ARBITRARILY CENSORED DATA SUBJECT TO COVARIATE MEASUREMENT ERROR

Xiaoyan Lin, University of South Carolina*
Lianming Wang, University of South Carolina

We develop a Bayesian estimation method based on the accelerated failure time (AFT) model for analyzing complicated survival data, in which the failure time may be exactly observed, left-, interval-, or right-censored and some of the covariates are subject to measurement errors. The distribution of the error term in the AFT model is assumed to be unknown and is modeled by the Dirichlet process normal mixture. An efficient Gibbs sampler is proposed for the posterior computation. The method is evaluated by a simulation study and is illustrated by an HIV data set.

e-mail: lin9@mailbox.sc.edu

SUBSAMPLE IGNORABLE MAXIMUM LIKELIHOOD FOR ACCELERATED FAILURE TIME MODELS WITH MISSING PREDICTORS

Nanhua Zhang, University of South Florida*
Roderick J. Little, University of Michigan

Missingness of predictors is common in survival analysis. In this paper, we review complete-case analysis and maximum likelihood estimation for accelarated failure time models with missing predictors, and propose a hybrid class, called subsample ignorable likelihood (SIL-AFT), which applies ignorable maximum likelihood method to a subsample of observation that are complete on one set of variables, but possibly incomplete on others. We give conditions on the missing data mechanism under which subsample ignorable likelihood method is consistent, while both complete-case analysis and maximum likelihood estimation mehtod are inconsistent. We illustrate the properties of the proposed method by simulation and apply the method to a real dataset.

e-mail: nzhang1@health.usf.edu

115. ENVIRONMENTAL AND ECOLOGICAL APPLICATIONS

THE EFFECT OF AIR POLLUTION CONTROL ON LIFE EXPECTANCY IN THE UNITED STATES: AN ANALYSIS OF 545 U.S. COUNTIES FOR THE PERIOD 2000 TO 2007

Andrew W. Correia, Harvard University*
Francesca Dominici, Harvard University

As a result of legislative efforts over the past three decades, there have been substantial and measureable improvements in ambient air quality in the United States. Over that same period, there have been continued improvements in population survival as well. While numerous epidemiological studies during this time have demonstrated that reductions in fine particulate matter air pollution (PM2.5) are associated with reductions in both cardiopulmonary and overall mortality, there has been substantially less work done to quantify the years of life gained from reductions in fine particulate air pollution. Here, we used a simple GEE approach, adjusting for temporal trends in other key predictors of morality, to directly assess the relationship between reductions in PM2.5 levels and changes in life expectancy in 545 U.S. counties for the period 2000 to 2007. Using stratified models, we additionally investigated the possibility of effect modification of several variables on the relationship between reductions in PM2.5 and life expectancy.

e-mail: acorreia@hsph.harvard.edu

MODELING SPACE-TIME QUANTILE SURFACES FOR NONSTATIONARY RANDOM FIELDS

Dana Sylvan, Hunter College of the City University of New York*

There is an increasing interest in studying time-varying quantiles, particularly for environmental processes. For instance, high pollution levels may cause severe respiratory problems, and large precipitation amounts can damage the environment, and have negative impacts on the society. Maximum readings would give a more relevant statistic for monitoring than average values. However, high order quantiles are preferred to maximum values, in order to increase statistical stability. In this paper, we use multidimensional kernel smoothing to map quantile fields in a large class of space-time processes, including non-Gaussian and non-stationary settings. We analyze theoretical properties, discuss implementation procedures, and illustrate the findings in simulation studies and applications to environmental data.

e-mail: dsylvan@hunter.cuny.edu

FAST COPULA-BASED SPATIAL REGRESSION FOR DISCRETE GEOSTATISTICAL DATA

John Hughes, University of Minnesota*

The copula-based model for geostatistical data offers compelling advantages over the more commonly used spatial generalized linear mixed model (SGLMM). But inference for copula-based models with discrete marginal distributions is challenging because the true likelihood is computationally intractable for all but the smallest datasets. In this talk I will develop the distributional transform (DT) approach to approximate classical and Bayesian inference for such models, and show the DT approach to be superior to competing approaches from both statistical and computational points of view. I will finish by comparing copula-based and SGLMM analyses of a real dataset from ecology.

e-mail: hughesj@umn.edu

MORTALITY EFFECTS OF PARTICULATE MATTER CONSTITUENTS IN A NATIONAL STUDY OF U.S. URBAN COMMUNITIES

Jenna R. Krall, Johns Hopkins Bloomberg School of Public Health*
Francesca Dominici, Harvard School of Public Health
Michelle L. Bell, Yale University
Roger D. Peng, Johns Hopkins Bloomberg School of Public Health

Variation in the chemical composition of total mass particulate matter less than 2.5 micrometers in diameter (PM2.5) has been suggested as a possible driver of the observed spatial and temporal differences in the effect of PM2.5 on mortality and morbidity. To date, national-level evidence concerning the mortality effects of PM2.5 constituents is limited and the effect of measurement error induced by spatial misalignment has not been comprehensively assessed. We estimated the association between all-cause

mortality and PM2.5 constituents in 72 urban communities across the United States for the years 2000-2005. We considered seven constituents that together compose 79-85% of PM2.5 mass: ammonium ion, elemental carbon, nitrate, organic carbon matter, silicon, sodium ion, and sulfate. National average, season-specific and region-specific effects were estimated for each constituent. Interquartile range increases in organic carbon matter and elemental carbon at a 1-day lag were associated with 0.60% (95% Posterior Interval [PI]: 0.09%,1.11%) and 0.75% (95% PI: 0.20%,1.30%) increases in mortality. We did not find that the mortality risks differed significantly between seasons or between regions. Furthermore, there is evidence that measurement error induced by spatial misalignment is an important factor to consider in estimating health risks of PM2.5 constituents.

e-mail: jkrall@jhsph.edu

ESTIMATING COVARIANCE PARAMETERS AND GENERALIZED LEAST SQUARES ESTIMATORS IN LINEAR MODELS WITH SPATIALLY MISALIGNED DATA

Kenneth K. Lopiano, University of Florida*
Linda J. Young, University of Florida
Carol A. Gotway, Centers for Disease Control

In environmental studies, relationships among variables that are misaligned in space are routinely assessed. Because the data are misaligned, kriging is often used to predict the covariate at the locations where the response is observed. Using kriging predictions to estimate regression parameters in linear regression models introduces both Berkson and classical measurement error. As a result, the Berkson error induces a covariance structure that is challenging to estimate. We characterize the measurement error as part of a broader class of Berkson error models and develop an estimated generalized least squares estimator using estimated covariance parameters. In working with the induced model, we fully account for the error structure and estimate the covariance parameters using restricted maximum likelihood and method of moments. We assess the performance of the estimators using simulation and illustrate the methodology using publicly available data from the Environmental Protection Agency. Finally, we extend the results to another change-of-support problem where the response is observed at the areal unit level and the covariate is observed at the point level using an example from the Centers for Disease Control's Environmental Public Health Tracking Program.

e-mail: klopiano@ufl.edu

COMPARING MAPS ACROSS TIME: SPATIO-TEMPORAL MORAN'S I IN STARMA MODELS

Nathan M. Holt, University of Florida*
Linda J. Young, University of Florida
Carol A. Gotway, Centers for Disease Control and Prevention

Generalizing the work of Moran (1950) and others, Cliff and Ord (1981) define Moran's I to be a ratio of quadratic forms in the regression errors of a simultaneous autoregressive (SAR) model. They then show that Moran's I may be employed to test for spatial independence among SAR model regression errors and, on this basis, Moran's I has since been identified as a measure of spatial clustering. Oftentimes interest lies in determining whether spatial clustering changes over time. One disadvantage of comparing Moran's I values across time is that apparently different spatial patterns may have the same value of Moran's I. As an example, for an outcome of a spatial process observed on a square lattice, the statistic is identical for all lattice rotations. In this work we define spatio-temporal Moran's I (STMI) to be a ratio of quadratic forms in spatio-temporal autoregressive moving average (STARMA) model regression errors. Using the proposed STMI, hypothesis tests of spatial, temporal, and spatio-temporal independence are developed. STMI tests are shown to be able to detect changes in pattern at two or more time points even when Moran's I is the same for each time.

e-mail: nateholt@ufl.edu

MODELING LOW-RANK SPATIALLY VARYING CROSS-COVARIANCES USING PREDICTIVE PROCESS WITH APPLICATION TO SOIL NUTRIENT DATA

Rajarshi Guhaniyogi, University of Minnesota*
Andrew O. Finley, Michigan State University
Rich Kobe, Michigan State University
Sudipto Banerjee, University of Minnesota

We extend earlier work on hierarchical multivariate spatial models to accommodate non-stationarity in the correlations among the outcomes as well as capturing the underlying spatial associations. Direct application of such multivariate models to even moderate-sized spatial datasets is often computationally infeasible because of the large number of parameters used to describe the nonstationary multivariate structures and cubic order matrix algorithms involved in estimation. Our methodological contribution comprises a new class of low-rank spatially-varying cross-covariance matrices that are non-degenerate and that effectively capture nonstationary covariances among the multiple outcomes. We provide theoretical and modeling insight into these constructions and elucidate certain implications of some common structural assumptions in building cross-covariance matrices. From a data analytic standpoint, we apply our methods to a soil nutrients dataset collected at La Selva Biological Station, Costa Rica. Here, interest lies in visual and statistical inference in the spatially-varying relationship among the residual spatial processes. Our framework produces substantive inferential tools such as maps of nonstationary cross-covariances that have hitherto not been easily available for environmental scientists and researchers.

e-mail: rajarshign84@gmail.com

116. NEXT GENERATION SEQUENCING

STATISTICAL MODELING OF CLOSELY LOCATED PROTEIN BINDING SITES USING PAIRED-END TAG (PET) ChIP-Seq DATA, WITH APPLICATION TO THE STUDY OF SIGMA70 FACTOR IN ESCHERICHIA COLI

Dongjun Chung, University of Wisconsin, Madison*
Jeff Grass, University of Wisconsin, Madison
Kevin Myers, University of Wisconsin, Madison
Patricia Kiley, University of Wisconsin, Madison
Robert Landick, University of Wisconsin, Madison
Sunduz Keles, University of Wisconsin, Madison

Chromatin immunoprecipitation followed by high throughput sequencing (ChIP-Seq) has revolutionized the study of gene regulation. In ChIP-Seq studies, paired-end tag (PET) technology has been used to reduce ambiguity in aligning tags to the reference genome and is considered as an important alternative to popular single-end tag (SET) technology, especially for the study of factors binding to repetitive regions. However, such advantages remain elusive in the study of prokaryotic genomes with few repetitive regions. This paper is motivated by the problem of identifying closely located sigma70 binding sites, which is a biologically important question in the study of Escherichia coli. We aimed to statistically assess PET and SET technologies from the view of improving resolution of binding site identification. We proposed a unified generative model for data that arise from SET and PET ChIP-Seq studies. We developed an Expectation-Conditional-Maximization algorithm based on this model. We showed with extensive simulation studies and a case study of sigma70 factor in E. coli, PET has clear advantages over SET in ChIP-Seq data for the study of proteins with complicated binding structures, such as closely located binding sites and unbalanced binding strengths.

e-mail: dchung4@wisc.edu

A GENERALIZED LINEAR MODEL FOR PEAK CALLING IN ChIP-Seq DATA

Jialin Xu, The Pennsylvania State University*
Yu Zhang, The Pennsylvania State University

Chromatin immunoprecipitation followed by massively parallel sequencing (ChIP-Seq) has become a routine for detecting genome-wide protein-DNA interaction. The success of ChIP-Seq data analysis highly depends on the quality of peak calling, i.e., to detect peaks of tag counts at a genomic location and evaluate if the peak corresponds to a real protein-DNA interaction event. The challenges in peak calling include 1) how to combine the forward and the reverse strand tag data to improve the power of peak calling, and 2) how to account for the variation of tag data observed across different genomic locations. We introduce a new peak calling method based on the generalized linear model (GLMNB) that utilizes negative binomial distribution to model the tag count data and account for the variation of background tags that may randomly bind to the DNA sequence at varying levels due to local genomic structures and sequence contents. We allow local shifting of peaks observed on the forward and the reverse stands, such that at each potential binding site, a binding profile representing the pattern of a real peak signal is fitted to best explain the observed tag data with maximum likelihood. Our method can also detect multiple peaks within a local region if there are multiple binding sites in the region.

e-mail: jxx120@psu.edu

A DYNAMIC SIGNAL PROFILE ALGORITHM COMBINED WITH A BAYESIAN HIDDEN ISING MODEL FOR ChIP-Seq DATA ANALYSIS

Qianxing Mo, Baylor College of Medicine*

Chromatin immunoprecipitation followed by next generation sequencing (ChIP-seq) is a powerful technique used in a wide range of biological studies. To systematically model ChIP-seq data, we construct a dynamic signal profile for each chromosome, and then model the profile using a fully Bayesian hidden Ising model. The proposed model naturally takes into account spatial dependency, global and local distributions of sequence tags. It can be used for one-sample and two-sample analyses. Through model diagnosis, the proposed method can detect falsely enriched regions caused by sequencing and/or mapping errors, which is usually not offered by the existing hypothesis-testing-based methods. The proposed method is illustrated using three transcription factor ChIP-seq data sets and four mixed ChIP-seq data sets, and compared with four popular and/or well-documented methods: MACS, CisGenome, BayesPeak and SISSRs. The results indicate that the proposed method achieves equal or higher sensitivity and spatial resolution in detecting transcription factor binding sites with a much lower false discovery rate.

e-mail: mo.quincy@yahoo.com

DETERMINING PROBABILITY OF RARE VARIANTS: DESIGN IMPLICATIONS FOR FAMILY-BASED SEQUENCING STUDIES

Wenyi Wang, University of Texas MD Anderson Cancer Center*
Gang Peng, University of Texas MD Anderson Cancer Center

There is an urgent need to find rare variants in order to better understand the genetic basis of human disease. Family-based sequencing studies have been performed, in search for functionally important genes. However, a high false discovery rate in calling rare variants continues to preclude a comprehensive downstream functional analysis across the genome. Recent studies have shown that linkage information among family members can help improve variant-calling accuracy. Existing methods are either post

hoc filters or only for family trios. In contrast, ongoing sequencing studies will include data on extended family members. We developed FamSeqPro, which computes the probability of variants in family-based sequencing data at the single base level. It updates the uncertainty measure of an individual's variant call by integrating family pedigree information with sequencing signal-to-noise ratios. We performed simulation studies to address study design questions with the objective of effectively sequencing for inherited genetic mutations among affected families. We also evaluated the performance of FamSeqPro, as compared to the single-individual-based method, in two datasets: previously published sequencing data with rare variants identified and Sanger data was used for validation; and the 1000 genomes project family trio sequencing data, which used HapMap dbSNP calls for validation.

e-mail: wwang7@mdanderson.org

A POWERFUL TEST FOR MULTIPLE RARE VARIANTS ASSOCIATION STUDIES THAT INCORPORATE SEQUENCING QUALITIES

Z. John Daye, University of Pennsylvania School of Medicine*
Hongzhe Li, University of Pennsylvania School of Medicine
Zhi Wei, New Jersey Institute of Technology

Next-generation sequencing data will soon become routinely available for association studies between complex traits and rare variants. Sequencing data, however, are characterized by the presence of sequencing errors. This makes it especially challenging to perform association studies of rare variants, which, due to their low minor allele frequencies, can be easily perturbed by genotype errors. Measures of sequencing qualities are generally available as quality scores for each individual genotype. But, despite the crucial role that sequencing qualities may play in the analysis of rare variants, they have, so far, been largely ignored in rare variants association studies. In this article, we propose the qMSAT, that allows the incorporation of sequencing qualities directly in association tests between complex traits and multiple rare variants. Simulation results based on quality scores from real data show that the qMSAT often dominates over current methods, that do not utilize quality information. In particular, the qMSAT can dramatically increase power over existing methods under moderate sample sizes and relatively low coverage. Moreover, in an application to the UCSD obesity data study, we identified using the qMSAT two functional regions (MGLL promoter and MGLL 3' untranslated region) where rare variants are associated with extreme obesity.

e-mail: zdaye@upenn.edu

117. NONPARAMETRIC METHODS

BOUNDED INFLUENCE NONLINEAR SIGNED-RANK REGRESSION

Huybrechts Frazier Bindele, Auburn University*

In this paper we consider weighted generalized-signed-rank estimators of nonlinear regression coefficients. The generalization allows us to include popular estimators such as the least squares and least absolute deviations estimators but by itself does not give bounded influence estimators. Adding weights results in estimators with bounded influence function. We establish conditions needed for the consistency and asymptotic normality of the proposed estimator and discuss how weight functions can be chosen to achieve bounded influence function of the estimator. Real life examples and Monte Carlo simulation experiments demonstrate the robustness and efficiency of the proposed estimator. An example shows that the weighted signed-rank estimator can be useful to detect outliers in nonlinear regression.

e-mail: achard4@gmail.com

ASYMPTOTIC OPTIMALITY AND EFFICIENT COMPUTATION OF THE LEAVE-SUBJECT-OUT CROSS-VALIDATION

Ganggang Xu, Texas A&M University, College Station*
Jianhua Huang, Texas A&M University, College Station

Although the leave-subject-out cross-validation (CV) has been widely used in practice for tuning parameter selection for various nonparametric and semiparametric models of longitudinal data, its theoretical property is unknown and solving the associated optimization problem is computationally expensive, especially when there are multiple tuning parameters. In this paper, by focusing on penalization methods, we show that the leave-subject-out CV is optimal in the sense that it is asymptotically equivalent to the empirical squared-error loss function. We also develop an efficient Newton-type algorithm to compute the penalty parameters that optimize the CV criterion. We use simulated and real data to demonstrate the effectiveness of the leave-subject-out CV in selecting both the penalty parameters and the working covariance matrix in generalized estimating equations.

e-mail: gang@stat.tamu.edu

■ = Student Award Winner | ★ = Presenter

UNCONDITIONAL TESTS TO MEASURE AGREEMENT FOR CATEGORIAL DATA IN APPLICATIONS TO A BRAIN TRAUMA STUDY

Guogen Shan, Brain Trauma Foundation*
Gregory Wilding, University at Buffalo
Changxing Ma, University at Buffalo
Alison Schonberger, Brain Trauma Foundation
Jamshid Ghajar, Brain Trauma Foundation

We consider an unconditional procedure to reduce the conservative of the exact conditional approach for measuring agreement between two tests whose outcome is either nominal or ordinal. The asymptotic approach is commonly used for large sample sizes, but may not be appropriate in the study with small sample sizes. The exact conditional approach was proposed to guarantee the nominal level of the test; however, it is often conservative in most statistical and medical applications. An alternative to measure the agreement is the unconditional approach, which is generally less conservative and more powerful than the conditional approach. We use a real example from a brain trauma study to illustrate the various test procedures. An extensive numerical study is provided to compare procedures and showed that the proposed unconditional approach has higher power as compared to competitors.

e-mail: gshan@braintrauma.org

GENERAL PIVOTAL GOODNESS OF FIT TEST BASED ON KERNEL DENSITY ESTIMATION

Hani M. Samawi, Georgia Southern University*
Robert Vogel, Georgia Southern University

In this paper we introduce a pivotal goodness of fit test based on empirical kernel density estimation. Our investigation reveals that the new test is more powerful than the traditional goodness of tests found in the literature; namely, the Chi-square and the Kolmogorov-Smirnov (KS) goodness of fit tests. Intensive simulation is conducted to examine the power of the proposed test. Data from a level I Trauma center are used to illustrate the procedures developed in this paper.

e-mail: hsamawi@georgiasouthern.edu

BERRY-ESSEEN-TYPE BOUNDS FOR GENERAL NONLINEAR STATISTICS, WITH APPLICATIONS TO PEARSON'S AND NON-CENTRAL STUDENT'S AND HOTELLING'S

Iosif Pinelis, Michigan Technological University*

Uniform and nonuniform Berry-Esseen (BE) bounds of optimal orders for general nonlinear statistics are obtained. Applications to Student's, Pearson's, and Hotelling's statistics are given, which appear to be the first known results of these kinds (with the exception of uniform BE bounds for Student's statistic). The proofs use a Stein-type method developed by Chen and Shao, a Cramer-type of tilt transform, exponential and Rosenthal-type inequalities for sums of random vectors established by Pinelis, Sakhanenko, and Utev, as well as a number of other, quite recent results motivated by this study. The method allows one to obtain bounds with explicit and rather moderate-size constants, at least as far as the uniform bounds are concerned.

e-mail: ipinelis@mtu.edu

BAYESIAN QUANTILE REGRESSION USING A MIXTURE OF POLYA TREES

Minzhao Liu, University of Florida*
Michael Daniels, University of Florida

Compared to mean regression, quantile regression is more robust and provides more complete information about the distribution. Traditional quantile regression methods include minimizing a target function by linear programming and Bayesian approaches using an asymmetric laplace distribution on the error term. However, these approaches are too restrictive with regards to the error distribution. In addition, it can be the case that quantile coefficients are not the same for different quantiles with heterogeneity in the variance of the error term contributing to this problem. To deal with these two problems, we propose a Bayesian quantile regression approach which assigns a mixture of Polya trees prior for the error distribution and allows for heterogeneity in the variability. The operating characteristics of this approach are examined through a wide array of simulations and the approach is applied to a real data set.

e-mail: liuminzhao@gmail.com

CONFIDENCE INTERVALS UNDER ORDER RESTRICTION

Yong Seok Park, University of Michigan*
John D. Kalbfleisch, University of Michigan
Jeremy MG Taylor, University of Michigan

In this paper, we consider the problem of constructing confidence intervals (CIs) for G independent normal population means under linear ordering constraints. For this problem CIs based on asymptotic distributions, likelihood ratio tests and bootstraps do not have good properties particularly when some of the population means are close to each other. We propose a new method based on defining intermediate random variables that are related to the original observations and using the CIs of the means of these intermediate random variables to restrict the original CIs from the separate groups. The coverage rates of the intervals are shown to exceed, but be close to, the nominal level for two groups, when the ratio of the variances is assumed known. Simulation studies show that the proposed CIs have coverage rates close to nominal levels with reduced average widths. An example of half-lives of an antibiotic is analyzed to illustrate the method.

e-mail: yongpark@umich.edu

118. SEMI-PARAMETRIC AND NON-PARAMETRIC MODELS

LOCALLY EFFICIENT ESTIMATION OF MARGINAL TREATMENT EFFECTS USING AUXILIARY COVARIATES IN RANDOMIZED TRIALS WITH CORRELATED OUTCOMES

Alisa J. Stephens, Harvard University*
Eric Tchetgen Tchetgen, Harvard University
Victor De Gruttola, Harvard University

Semiparametric methods have been developed to increase efficiency of inferences in randomized trials by incorporating baseline covariates. Current literature demonstrates locally efficient estimators of marginal treatment effects when outcomes are independent. We derive semiparametric locally efficient estimators of marginal mean treatment effects when outcomes are correlated, which occurs in clinical trials with clustered or repeated-measures data. The resulting estimating equations modify existing generalized estimating equations (GEE) by identifying the efficient score under a mean model for marginal effects when data contain baseline covariates. Locally efficient estimators are implemented for clustered and longitudinal data with continuous outcomes, and clustered data with binary outcomes. Methods are illustrated through application to the AIDS Clinical Trial Group Study #398, a longitudinal randomized clinical trial that compared the effects of various protease inhibitors in HIV-positive subjects with antiretroviral therapy failure. The new estimators are compared to various existing estimators through simulation and data analysis.

e-mail: astephen@hsph.harvard.edu

KERNEL MACHINE QUANTILE REGRESSION OF MULTI-DIMENSIONAL GENETIC DATA

Dehan Kong, North Carolina State University*
Arnab Maity, North Carolina State University
Jung-Ying Tzeng, North Carolina State University

We consider quantile regression for partially linear models where an outcome of interest is related to covariates and a genetic pathway with the covariate effects being modeled parametrically and the pathway effect of multiple genetic variates modeled using kernel machines. We propose a fast and efficient algorithm to solve the corresponding optimization problem and also introduce a powerful test for detecting the overall gene pathway effect. Our test is motivated by traditional score test, and borrows the idea of permutation test. We show via simulation study that the proposed test is much more powerful than standard tests such as Wald test and rank test. We evaluate our estimation and testing procedures using simulation study.

e-mail: dkong2@ncsu.edu

AN IMPROVED METHOD FOR CHOOSING THE SMOOTHING PARAMETER IN A SEMI-PARAMETRIC CHANGE-POINT MODEL

Sung Won Han, University of Pennsylvania*
Theresa Busch, University of Pennsylvania
Mary Putt, University of Pennsylvania

In an animal model, the extent and duration of the reduction in blood flow in solid tumors appears to be a key determinant of subsequent tumor response to therapy. Estimating the change-points corresponding to the initial reduction and the subsequent stabilization of flow is challenging because the baseline blood flow is not easily fit to a parametric model. We modeled the data using a smoothing spline for the baseline curvature and a parametric component to add a linear decrease in flow to the baseline between the change-points. While a generalized cross validation (GCV) is commonly used as a a criteria for choosing the smoothing parameter in similar "partial spline" models, simulation indicates that the resulting estimates of the blood flow at the change-points have substantial bias and variance. We observed that GCV leads to under-smoothing of the data particularly with larger curvature in the baseline flow. We propose a modification to GCV that depends on the change-size to noise ratio and that corrects for this tendency to under-smooth. Results from both simulation and data collected in recent experiments suggest that this new method yields substantial improvement in the bias and variance of the resulting estimates.

e-mail: hansungw@mail.med.upenn.edu

SEMIPARAMETRIC BAYESIAN JOINT MODELING OF A BINARY AND CONTINUOUS OUTCOME

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

In dose-response model, many studies collect toxicity data on correlated outcomes. For example, fetal weight and malformation in developmental toxicology studies are measured on the same live fetuses. To analyze these outcomes, joint modeling can result in more efficient inferences than independent models (Regan and Catalano, 1999). Most methods for joint modeling have assumed standard parametric response distributions for both frequentist and Bayesian approaches (Catalano and Ryan, 1992; Dunson, 2000). However, it is possible that toxicity responses vary in the location and shape of distribution with dose, which may not be easily captured by standard parametric distributions. Dunson and Park (2008) proposed a kernel stick-breaking process (KSBP), which allows location and shape to change with dose. We propose a semiparametric Bayesian joint model for a binary and continuous response. In our model, KSBP prior is assigned to distribution of a random effect shared across outcomes, which allows flexible changes in shape shared across outcomes and different effects on location with dose. The proposed model provides more accurate estimates of potency when the data does not satisfy assumptions of parametric models. We evaluate our approach using extensive simulation data. We apply our method to the DDT metabolite DDE data.

e-mail: hwang.176@osu.edu

SEMIPARAMETRIC SINGLE INDEX INTERACTION MODEL IN 1-M MATCHED CASE-CROSSOVER STUDIES

Chongrui Yu, Virginia Polytechnic Institute and State University*
Inyoung Kim, Virginia Polytechnic Institute and State University

Single-index models have been used in many applications such as biometrics and economics, where multidimensional regression models are often encountered. However, since single index model assumes a nonparametric regression function of linear combination among variables, it can not take into account of interactions among variables. Therefore, in this paper, we propose a semiparametric single index interaction model to automatically detect interactions among variables in a matched case-crossover study. Regression splines are used to model nonparametric function of the "index", which is a combination of variables and their interactions. We develop both frequentist and Bayesian approaches to fit the model. Two Bayesian methods are developed using two different priors: the prior distribution of index is based on (1) Fisher-von Mises distribution and (2) polar coordinates representation. Based on these two types of priors, a number of mixture priors are developed to detect variables as well as their interaction selection in the low order interaction models. Simulation results indicate our Bayesian methods provide more significant improvement than a frequentist method. We demonstrate our approaches using an epidemiological example of a 1-4 bi-directional case-crossover study.

e-mail: yucr@vt.edu

GENERALIZED METHOD OF WEIGHTED MOMENTS: A ROBUST ESTIMATOR OF POLYTOMOUS LOGISTIC MODEL

Xiaoshan Wang, University of North Carolina at Chapel Hill*
Pranab K. Sen, University of North Carolina at Chapel Hill

The maximum likelihood estimation (MLE) method, typically used for polytomous logistic regression, is prone to bias due to both misclassification in outcome and contamination in the design matrix. Hence, a robust estimator is needed. In this study, we propose a robust method for nominal response data with continuous covariates. A generalized method of weighted moments (GMWM) approach is developed for dealing with contaminated polytomous response data. In this approach, distances are calculated based on individual sample moments. And Huber weights are applied to those observations with large distances. Mellow-type weights are also used to downplay leverage points. We describe theoretical properties of the proposed approach. Simulations suggest that the GMWM performs very well to correct contamination-caused biases. An empirical application of the GMWM estimator on a survey data demonstrates its usefulness.

e-mail: xiwang@bios.unc.edu

■ = Student Award Winner | * = Presenter