Southern Regional Council on Statistics (SRCOS) Summer Research Conference

Hickory Knob State Resort
McCormick, South Carolina
June 5–8, 2011

Abstracts of Talks
(In Order of Presentation)

Session I
Time: June 6, Monday, 8:30-10:30AM
Topic: Multivariate Methods and Censored Regression
Organizer: KB Kulasekera, Clemson University

- 8:30–9:05
  Statistical Inference on the High-Dimensional Covariance Matrix
  Xiaoqian Sun [xsun@clemson.edu]
  Department of Mathematical Sciences, Clemson University, Clemson, SC
  Abstract: Statistical inference on the covariance matrix is one fundamental problem in multivariate analysis. Statistical theory of estimation and hypothesis testing on the Gaussian covariance matrix has well been developed in the classical setting when the sample size is larger than the number of variables. However, many applications of modern multivariate statistics involve a large number of variables and hence a large covariance matrix. In many situations such as microarray data, the number of variables is much larger than the sample size and thus a lot of classical procedures such as the likelihood ratio test will fail due to the singularity of the sample covariance matrix. In this talk, I will review some existing strategies of hypothesis testing on the high-dimensional Gaussian covariance matrix in the literature and then propose some new approaches for testing if the covariance matrix is proportional or equal to the identity matrix. Asymptotic results of the proposed test statistics are established under the setting when both the sample size and the number of variables go to infinity. In addition, some strategies on constructing a Stein-type shrinkage estimator of the high-dimensional covariance matrix will also be discussed. (Joint work with Thomas Fisher and Colin Gallagher)

- 9:05-9:40
  Modeling and Forecasting Functional Time Series
  Lily Wang [lilywang@uga.edu]
  Department of Statistics, University of Georgia, Athens, GA
  Abstract: A novel method is proposed for forecasting a time series of smooth curves, using functional principal component analysis in combination with time series modeling and scores forecasting. We achieve the smoothing, dimension reduction and prediction at the same time with the expedient computation. The work is motivated by the demand to forecast the time series of economic functions, such as Treasury bond yield curves. Extensive simulation studies have been carried out to compare the prediction accuracy of our method with other competitor's methods. The proposed methodology is applied to forecasting the yield curves of UK government bond.

- 9:40–10:15
  Local linear regression with censored data
  Meng Zhao [mzhao@math.msstate.edu]
  Department of Mathematics and Statistics, Mississippi State University, Starkville, MS
  Abstract: We propose a self-consistency based method to conduct local linear estimation of mean functions for general regression models with censored data. Self-consistent estimation techniques are frequently used to obtain consistent estimation of survival functions when data are censored. Such techniques have also been proven useful for estimation of linear regression parameters in censored setting. We extend such methods to local linear estimation of the regression functions for nonparametric models. The bandwidth will be selected adaptively according to the data. Simulation studies are conducted. Some basic asymptotic results are also obtained.

Session II
Time: June 6, Monday, 10:45AM–12:45PM
Topic: New Directions in Clinical Trials
Organizer: Michael Kutner, Emory University, Atlanta, GA

- 10:45–11:20
  Clinical Trials and Personalized Medicine
  Michael Kosorok [kosorok@unc.edu]
  Department of Biostatistics, University of North Carolina, Chapel Hill, NC
  Abstract: In this presentation, we discuss contemporary research issues in how clinical trials can be better utilized to discover and evaluate treatments. We describe the Innovative Methods Program for Advancing Clinical Trials (IMPACT), a new collaborative program for developing new methods for clinical trials. The initial funding mechanism for IMPACT is a large program project (P01) grant from the National Cancer Institute. The program is a collaboration between Duke, North Carolina State University, and the University of North Carolina at Chapel Hill, led by statisticians but including clinical scientists, biologists and other experts. In this talk, and in the two other talks of this session, we will give some of the flavor of the topics addressed in the program. For the first talk, we highlight some new ideas for discovering and validating personalized treatment regimens in cancer, with specific examples in lung and colorectal cancer. We describe fundamentally new paradigms that allow clinical trials to be used for treatment discovery in addition to the more traditional purposes.

- 11:20-11:55
  Innovative Clinical Trial Designs
  Sin-Ho Jung and Stephen George [stephen.george@duke.edu]
  Department of Biostatistics, Duke University, Durham, NC
  Abstract: Innovative clinical trial designs can potentially require fewer patients, save resources, and accelerate cancer drug development. Although much effort has been put into analysis methods for complicated data structures, the design aspect has not kept pace. One of the IMPACT projects is concerned with improving the design of clinical trials in three broad areas: using joint models for longitudinal and survival data; group randomized cancer prevention trials with survival and recurrent event outcomes; and cancer drug development.

In this presentation I will focus on two areas under the latter topic: targeted clinical trials and predictive phase II trials. In addition, I will describe the data acquisition core of the IMPACT program.

- 11:55–12:30
  Improving Efficiency of Inferences on Treatment Effect and Identifying Optimal Treatment Strategies
  Marie Davidian [davidian@ncsu.edu]
  Department of Statistics, North Carolina State University, Raleigh, NC
  Abstract: The P01 Program Project grant underlying IMPACT involves five integrated research projects. In this talk, we provide an overview of ongoing research addressing specific aims for two of these projects. The first research topic involves a study of the use of modern variable selection methods for identifying baseline covariates to be incorporated in covariate-adjusted inferences on treatment effect in traditional clinical trials. The second focuses on evaluation of the relative merits of two different reinforcement learning methods for identifying the optimal strategy for using individual patient information to determine treatment. We conclude by discussing plans for dissemination of public use software implementing methods developed in all five projects.

Session III

Time: June 6, Monday, 7:15PM–9:15PM

Topic: Panel Discussion on Effective Teaching of Introductory Statistics/Biostatistics Courses

Organizer: Michael Kutner, Emory University, Atlanta, GA

- Panelist 1:
  Teaching the BIG IDEAS of Statistics in the Introductory Course
  Christine Franklin [chris@stat.uga.edu]
  Department of Statistics, University of Georgia, Athens, GA
  Abstract: For so many students, we only have one opportunity to help the student develop statistical reasoning skills - the college intro course. How do we get the student to buy into the importance of statistical thinking during the first week of class and illustrate for the student the "big picture" of statistical reasoning tying together the process of first asking a statistical question of interest, designing an appropriate study for collecting data, exploring the data, and then after analyzing the data, answering the question of interest. Too often, the intro course is taught as a laundry list of disjointed topics and students rarely make these connections. I will present an example of using a case study to help students (through hands on simulation) understand during the first week of class the connection of the "big ideas" in statistics. I will also briefly discuss my philosophy of teaching the introductory course.

- Panelist 2:
  Teaching Introductory Statistics in the Information Age
  Patrick Kilgo [pkilgo@sph.emory.edu]
  Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA
  Abstract: A generation ago, it was perfectly appropriate for a professor to serve as the chief imparter of knowledge since there was not wide availability to learning resources and access

to information was often limited or inconvenient. However, students today have information readily available at their fingertips with just a few clicks of a computer mouse and college libraries have many types of user-friendly media. Problem-based learning (PBL) tries to capitalize on these advances by combining the wide access to information with a very basic principle of learning - that we learn best actively by performing tasks rather than by passively listening. The foundation of PBL is that learning is best achieved in the context of solving a problem. For the past two academic years, we have successfully incorporated PBL into some of our large introductory biostatistics graduate courses. The purpose of my talk will be to describe both our implementation of PBL methods and our experience with training students using this approach.

- Panelist 3:
  Introductory Statistics Courses for the Modern Day Masses
  Azhar Nizam [anizam@sph.emory.edu]
  Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA
  Abstract: Growing numbers of students are using statistical methods at the undergraduate and graduate school levels, and growing numbers of employees in all sectors encounter and must deal with data in various forms in their work lives. Many of these people will not take any formal courses dedicated to applied statistical methods. Some will take a single course, in high school or college. Relatively few non-statistics majors will take additional courses in statistics. Given that an introductory statistics course is likely to be the first and last formal statistics course that a student will take, it is important to re-examine objectives for these courses, to determine where they need to be tailored to meet the needs of the modern student, and to determine how they can be tailored without sacrificing rigor. I will review the evolution of objectives in introductory courses that I have taught over the past 20 years, and will discuss challenges faced by modern day students and teachers in these courses.

Session IV
Time: June 7, Tuesday, 8:30-10:30AM
Topic: Statistical Genetics
Organizer: Yong Chen, University of Texas School of Public Health, Houston, TX

- 8:30–9:15
  Genetics Studies of Multivariate Traits
  Heping Zhang [Heping.Zhang@Yale.EDU]
  Department of Biostatistics, Yale University, New Haven, CT
  Abstract: Identifying the risk factors for comorbidity is important in psychiatric research. Empirically, studies have shown that testing multiple, correlated traits simultaneously is more powerful than testing a single trait at a time in association analysis. Furthermore, for complex diseases, especially mental illnesses and behavioral disorders, the traits are often recorded in ordinal scales. In absence of covariates, nonparametric association tests have been developed for multiple (ordinal and/or quantitative) traits to study comorbidity. However, genetic studies generally contain measurements of some covariates that may confound the relationship between the risk factors of major interest (such as genes) and the outcomes. While it is relatively straightforward to include the covariates in the analysis of multiple

quantitative traits, it is challenging for multiple ordinal traits. In this article, we propose a weighted test statistic based on a generalized Kendall's tau to adjust for the effects of the covariates. We conducted simulation studies to compare the type I error and power of our proposed test with an existing test. The empirical results suggest that our proposed test increases the power of testing association when adjusting for the covariates. We further demonstrate the advantage of our test by analyzing a data set on genetics of alcoholism.

This presents a series of joint work with Ching-Ti Liu, Wensheng Zhu, Yuan Jiang, and Xueqin Wang.

- 9:15-9:45
  Network-based Bayesian Variable Selection Approach to Genome-wide Association Studies Data
  Peng Wei [Peng.Wei@uth.tmc.edu]
  Division of Biostatistics, University of Texas School of Public Health, Houston, TX
  Abstract: Genome-wide association studies (GWAS) have rapidly become a standard method for disease gene discovery. To overcome the limitations of single-SNP analysis, there have been increasing efforts recently on GWAS pathway analysis, aiming at combining SNPs with moderate signals. However, a major drawback of current pathway-based methods is that interactions among genes within a pathway are ignored and genes are treated as exchangeable, leading to inefficient use of biological prior knowledge and loss of power. Here we propose a flexible Bayesian nested mixture model to incorporate genome-wide gene-gene interaction information embedded in gene networks into gene-based analysis of GWAS data. We carry out parameter estimation and inference based on MCMC samples in a Bayesian variable selection framework. Applications to real GWAS datasets, together with simulation studies, demonstrates the extra power gained by integrating gene networks with GWAS data.

- 9:45–10:15
  Assessing Genetic Association in Case-Control Studies with Unmeasured Population Substructure
  Yong Chen [Yong.Chen@uth.tmc.edu]
  Division of Biostatistics, University of Texas School of Public Health, Houston, TX
  Abstract: The case-control study design is one of the main tools for detecting associations between genetic markers and disease. It is well known population substructure (PS) can lead to spurious association between disease status and a genetic marker if the prevalence of disease and the marker allele frequency vary across subpopulations. In this talk, we proposed a novel statistical method to estimate the association in case-control studies with potential population substructure. The proposed method takes two steps. First, the information on genomic markers and disease status is used to infer population substructure; second, the association between disease and any one marker adjusting for the population substructure is then modeled and estimated parametrically through polytomous logistic regression. The performance of the proposed method, relative to others, on bias, coverage probability and computational time, is assessed through simulations. Finally, this method is applied to an end-stage renal disease study in African Americans population.

Session V
Time: June 7, Tuesday, 10:45AM–12:45PM

Topic: Analysis of Large and High-Dimensional Data
Organizer: Wenguang Sun, North Carolina State University, Raleigh, NC

- 10:45–11:30
  Paradigm Free Mapping with Sparse Regression
  Ian Dryden [dryden@mailbox.sc.edu]
  Department of Statistics, University of South Carolina, Columbia, SC
  Abstract: The ability to detect single trial responses in functional MRI (fMRI) studies is important, yet traditionally a known paradigm is used where the timing of activations has to be specified. The statistical task in our work is to detect signals in space and time in very high-dimensional datasets, taking into account the physical properties inherent in the data collection. Paradigm Free Mapping (PFM) is a method that detects single trial responses without specifying prior information on the timing of the events. The PFM method is based on the deconvolution of the fMRI signal using a hemodynamic response function, and involves a ridge regression estimator for signal deconvolution and a baseline signal period for statistical inference. A sparse version of PFM uses the Dantzig Selector, and this method obtains high detection rates of activation, comparable to a model-based analysis, but requiring no information on the timing of the events or baseline period. The practical operation of this sparse paradigm free mapping method was assessed with single-trial fMRI data acquired at 7 Tesla, where it automatically detected all task-related events. The work is joint with Cesar Caballero Gaudes, Natalia Petridou, Susan Francis and Penny Gowland.

- 11:30-12:00
  Multiple Testing for Pattern Identification, with Applications to Microarray Time-Course Experiments
  Zhi Wei [zhiwei@njit.edu]
  New Jersey Institute of Technology, Newark, NJ
  Abstract: In time-course experiments, it is often desirable to identify genes that exhibit a specific pattern of differential expression over time and thus gain insights into the mechanisms of the underlying biological processes. Two challenging issues in the pattern identification problem are: (i) how to combine the simultaneous inferences across multiple time points and (ii) how to control the multiplicity while accounting for the strong dependence. We formulate a compound decision-theoretic framework for set-wise multiple testing and propose a data-driven procedure that aims to minimize the missed set rate (MSR) subject to a constraint on the false set rate (FSR). A hidden Markov model (HMM) is generalized to capture the temporal correlation in the gene expression data. Both theoretical and numerical results are presented to show that our data-driven procedure controls the multiplicity, provides an optimal way of combining simultaneous inferences across multiple time points, and greatly improves the conventional combined p-value methods. In particular, we demonstrate our method in an application to a study of systemic inflammation in humans for detecting early and late response genes.

- 12:00–12:30
  Shrinkage Estimators for Out-of-Sample Prediction in High-Dimensional Linear Models
  Lee Dicker [ldicker@stat.rutgers.edu]
  Department of Statistics and Biostatistics, Rutgers University, Pistcataway, NJ
  Abstract: We discuss the out-of-sample prediction error (predictive risk) associated with two classes of shrinkage estimators for the linear model: James-Stein type shrinkage estimators

and ridge regression estimators. Our study is motivated by problems in high-dimensional data analysis and our results are especially relevant to settings where both the number of predictors and observations are large. Moreover, our results provide a means for a detailed, yet transparent comparative analysis of the different estimators, which helps to shed light on their relative merits. For instance, we utilize results from random matrix theory to obtain explicit closed form expressions for the asymptotic predictive risk of the estimators considered herein (in fact, many of the relevant results are non-asymptotic). Additionally, we identify minimax ridge and James-Stein estimators, which outperform previously proposed shrinkage estimators, and prove that if the population predictor covariance is known - or if an operator norm-consistent estimator for the population predictor covariance is available - then the ridge estimator has smaller predictive risk than the James-Stein estimator.

Banquet Talk
Recollections and Reflections
Michael Kutner [mkutner@emory.edu]
Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA
Abstract: After having spent over 50 professional years as a mathematician, biostatistician, professor, author, and administrator, I will offer some recollections of my experiences that turned out to be either good or bad or ugly. I'll then offer some reflections. Stay tuned, you may not want to miss this "historic" event.

Session VI
Time: June 8, Wednesday, 8:30-10:30AM
Topic: Measurement Error and Latent Modeling
Organizer: Joshua Tebbs, University of South Carolina, Columbia, SC

- 8:30–9:05
  On Measurement Error Problems with Predictors Derived from Stationary Stochastic Processes and Application to Cocaine Dependence Treatment Data
  Yehua Li [yehuali@uga.edu]
  Department of Statistics, University of Georgia, Athens, GA
  Abstract: In a cocaine dependence treatment study, we use linear and nonlinear regression models to model post-treatment cocaine craving scores and first cocaine relapse time. A subset of the covariates are summary statistics derived from baseline daily cocaine use trajectories, such as baseline cocaine use frequency and average daily use amount. These summary statistics are subject to estimation error and can therefore cause biased estimators for the regression coefficients. Unlike classical measurement error problems, the error we encounter here is heteroscedastic with unknown distribution, and there are no replicates for these covariates or instrumental variables. We propose two robust methods to correct for the bias: a computationally efficient method-of-moment bias-correction method for linear regression models and a subsampling extrapolation method that is generally applicable to both linear and nonlinear regression models. Simulations and an application to the cocaine dependence treatment data are used to illustrate the efficacy of the proposed methods.

- 9:05-9:40
  Informative Model Specification Tests Using Coarsened Data

Xianzheng (Shan) Huang [huang@stat.sc.edu]
Department of Statistics, University of South Carolina, Columbia, SC
Abstract: Inappropriate model assumptions can result in erroneous statistical inference. Adverse effects of model misspecification on statistical inference are usually studied and deemed detectable only in the unrealistic scenario where the true model is known. We propose a simple idea that utilizes coarsened data to reveal the existence of bias in inference due to model misspecification without the knowledge of the truth. This idea leads to a versatile class of model diagnostic methods, which can disentangle multiple sources of misspecification coexisting in, for instance, a mixed model. Compared to the existing diagnostic methods, the new methods are more informative in that they can pinpoint the source(s) of misspecification and thus direct model correction. A key ingredient of these methods is to design a coarsening mechanism strategically and to study the interaction between the mechanism and model misspecification. I will illustrate these mainly in the context of mixed models.

- 9:40–10:15
Testing for the effect of a genetic pathway in longitudinal/clustered data using kernel machine regression
Arnab Maity [amaity@ncsu.edu]
Department of Statistics, North Carolina State University, Raleigh, NC
Abstract: There is a growing scientific interest to test for genetic effects on disease by considering a set of genes that may be on the same biological pathway. Genes within a pathway may interact in functional ways to influence the progression of disease. Kernel machine regression has been introduced as a way to model a pathway effect, either parametrically or nonparametrically. We consider kernel machine regression for the testing of a genetic pathway effect in the longitudinal/clustered data setting, where a continuous disease outcome is measured repeatedly, possibly over time, for each subject. We develop a score-based test statistic for testing the effect of the genetic pathway accounting for the within subject correlation in the outcome variable. In addition, we present a simulation study to investigate the power of the test for different correlation structures, and we compare its performance with the test without accounting for correlation.

Session VII
Time: June 8, Wednesday, 10:45AM–12:45PM
Topic: Time Ordered Data
Organizer: Robert Lund, Clemson University

- 10:45–11:30
A New Method of Modeling Integer Count Time Series
Robert Lund [Lund@clemson.edu]
Department of Mathematical Sciences, Clemson University, Clemson, SC
Abstract: This talk proposes a new but simple method of modeling stationary time series of integer counts. Previous work has focused on thinning methods and classical time series autoregressive moving-average difference equations; in contrast, our methods use a renewal process to generate a correlated sequence of Bernoulli trials. By superpositioning independent copies of such processes, stationary series with binomial, Poisson, geometric, or any

other discrete marginal distribution can be readily constructed. The model class proposed is parsimonious, non-Markov, and readily generates series with either short or long memory autocovariances. The model can be fitted with linear prediction techniques for stationary series. As an example, a stationary series with binomial marginal distributions is fitted to the number of rainy days in 210 consecutive weeks at Key West, Florida.

- 11:30-12:00
  A Bayesian Approach to Seasonal Adjustment of Long Memory Time Series
  Scott Holan [HolanS@missouri.edu]
  Department of Statistics, University of Missouri, Columbia, MO
  Abstract: Research into long memory processes has recently spread to the modeling of seasonality through the use of generalized exponential (GEXP) time series models. This talk describes the GEXP model and introduces the new Seasonal Fractional Exponential (SFEXP) model. We explore the fit of these models to economic time series data and present an application of seasonal long memory modeling to the problem of seasonal adjustment. In particular, we discuss a structural approach to obtaining component models for seasonal and trend in the context of long memory, and use these models to obtain minimum mean square error signal extraction estimates. The approach we propose is fully Bayesian and thus naturally quantifies the uncertainty in the signal extraction estimates. Finally, this technique is illustrated on several economic time series.

- 12:00–12:30
  Renewal Processes with Periodic Dynamics
  Brian Fralix [bfralix@clemson.edu]
  Department of Mathematical Sciences, Clemson University, Clemson, SC
  Abstract: This talk focuses on a periodic discrete renewal process, where the distribution of the length between two points is dependent on the position of the left endpoint. For this type of point process, we derive closed-form expressions for the limiting distribution of the corresponding age process.