Program

The 3rd Workshop on Biostatistics and Bioinformatics:
Celebrating the World of Statistics

Department of Mathematics and Statistics
Georgia State University
May 9-11, 2014
Contents

Sponsor ................................................................................................................................. 3
Organizing Committee ........................................................................................................... 3
Keynote and Invited Speakers ............................................................................................. 3
Acknowledgements ................................................................................................................ 4
Conference Schedule .......................................................................................................... 5
Keynote and Invited Talks .................................................................................................... 10
Poster Abstracts .................................................................................................................... 20
Sponsor

The National Science Foundation (NSF), Institute of Mathematical Statistics (IMS), International Chinese Statistical Association (ICSA), The Georgia State University Research Foundation, and the Department of Mathematics and Statistics in the Georgia State University

Organizing Committee

Yichuan Zhao (Chair), Niharika Muriki, Xin Qi

Department of Mathematics and Statistics

Georgia State University

Keynote Speaker

Xihong Lin, Harvard University

Invited Speaker

Yang Feng, Columbia University

Varghese George, Georgia Regents University

Jiashun Jin, Carnegie Mellon University

Jian Kang, Emory University

Soumendra Lahiri, North Carolina State University

Yi Li, University of Michigan

Ping Ma, University of Georgia

Edsel Pena, University of South Carolina
Annie Qu, University of Illinois

Glen Satten, CDC, Atlanta

Debajyoti Sinha, Florida State University

Peter Song, University of Michigan

T. N. Sriram, University of Georgia

Tony Sun, University of Missouri

Hongkun Wang, Georgetown University

Yichao Wu, North Carolina State University

Hongyan Xu, Georgia Regents University

Acknowledgements

The organizers thank Earnestine Collier-Jones, Sandra Ahuama-Jonas, Yvonne Pierce, Leslie Meadows, Tu Tran, and other volunteers for their great efforts in setting up this workshop and making it run successfully.
Conference Schedule

All workshop sessions meet in the room 150, College of Education Building, 30 Pryor Street, Atlanta, GA 30303.

Friday, May 9, 2014
2:00-6:00 pm  Registration: on the 7th Floor, College of Education Building, 30 Pryor Street.

Saturday, May 10, 2014
8:00-8:35 am  Registration: on the 1st Floor, College of Education Building, 30 Pryor Street.

8:35-8:45 am  Conference Welcome: Guantao Chen, Chair of the dept., Georgia State University
Opening Remarks:  Mark Becker, President, Georgia State University

8:45-9:45 am  Session 1 (Keynote Talk): Chair: Yichuan Zhao, Georgia State University

Statistical Genetics and Genomics in the Big Data Era: Opportunities and Challenges in Research and Training
Xihong Lin, Harvard University

9:45-10:00 am  Break: Refreshments
10:00-11:50 am  **Session 2: Chair: Xin Qi, Georgia State University**

*Empirical Likelihood in high dimensions*

Soumendra Lahiri, North Carolina State University

*Classes of Multiple Decision Functions Controlling FWER and FDR*

Edsel Pena, University of South Carolina

*Leveraging in Big Data Regression*

Ping Ma, University of Georgia

11:50-2:00 pm  **Lunch Time**

2:00-3:45 pm  **Session 3: Chair: Yang Feng, Columbia University**

*Modeling time-varying effects for high-dimensional covariates: a new Gateaux-differential boosting approach*

Yi Li, University of Michigan

*Spectral Clustering*

Jiashun Jin, Carnegie Mellon University

*Bayesian Spatial Variable Selection for Ultra-High Dimensional Neuroimaging Data: A Multiresolution Approach*

Jian Kang, Emory University
3:45-4:00 pm  Break: Refreshments

4:00-5:40 pm  **Session 4: Chair:** Ruiyan Luo, Georgia State University

  *Cost and Cost-Effectiveness Analysis with Incomplete Data*

  Hongkun Wang, Georgetown University

  *Simple estimation procedures for regression analysis of interval-censored failure time data under the proportional hazards model*

  Tony Sun, University of Missouri

  *Safe Methods and Sample Size Determination for Non-Inferiority Trial with Survival Response*

  Debajyoti Sinha, Florida State University

5:40-5:55 pm  Break: Refreshments

5:55-6:55 pm  **Poster Session: Chair:** Remus Osan, Georgia State University:

  Room 150, College of Education Building

7:20-9:30 pm  **Workshop Banquet:** the Sun Dial Restaurant (Westin Peachtree Plaza): 210 Peachtree Street NW, Atlanta, GA 30303.
All workshop sessions meet in the room 150, College of Education Building

Sunday, May 11, 2014

8:00-8:30 am  **Registration**: on the 1st Floor, College of Education Building, 30 Pryor Street.

8:30-10:15 am  **Session 5**:  **Chair**: Jiashun Jin, Carnegie Mellon University

  *Regression analysis of networked data*
  *Peter Song, University of Michigan*

  *Sufficient Dimension Reduction in Binary Classification*
  *Yichao Wu, North Carolina State University*

  *Feature Augmentation via Nonparametrics and Selection (FANS) in High Dimensional Classification*
  *Yang Feng, Columbia University*

10:15-10:30 am  **Break**: Refreshments

10:30-12:05 pm  **Session 6**:  **Chair**: Peter Song, University of Michigan

  *Testing Association without Calling Genotypes Allows for Systematic Differences in Read Depth between Cases and Controls*
  *Glen Satten, CDC, Atlanta*

  *Determination of sample size for a multi-class classifier based on single nucleotide polymorphisms: A Volume Under the Surface approach*
  *T. N. Sriram, University of Georgia*
Methods for Differential Methylation Analysis with Next Generation Sequencing

Hongyan Xu, Georgia Regents University

12:05-12:10pm Final Remarks by Yichuan Zhao
Keynote Talk

Statistical Genetics and Genomics in the Big Data Era:
Opportunities and Challenges in Research and Training

Xihong Lin

Department of Biostatistics
Harvard School of Public Health

The human genome project in conjunction with the rapid advance of high throughput technology has transformed the landscape of health science research. The genetic and genomic era provides an unprecedented promise of understanding genetic underpinnings of complex diseases or traits, studying gene-environment interactions, predicting disease risk, and improving prevention and intervention, and advancing personalized medicine. A large number of genome-wide association studies conducted in the last ten years have identified over 1,000 common genetic variants that are associated with many complex diseases and traits. Massive next generation sequencing data as well as different types of ’omics data have become rapidly available in the last few years. These big genetic and genomic data present statisticians with many exciting opportunities as well as challenges in data analysis and interpretation of results. They also call for more interdisciplinary knowledge and research, e.g., in statistics, machine learning, data curation, molecular biology, genetic epidemiology and clinical science. In this talk, I will discuss some of these challenges, such as analysis of next generation sequencing association studies; integrative genomics, which integrates different types of ’omics data; and reproducible research. I will also discuss strategies of training next generation quantitative genomic scientists at the interface of statistical genetics and genomics, computational biology and genetic epidemiology, to meet these challenges.
Invited Talks

Feature Augmentation via Nonparametrics and Selection (FANS) in High Dimensional Classification

Yang Feng
Columbia University

We propose a high dimensional classification method that involves nonparametric feature augmentation. Knowing that marginal density ratios are building blocks of the Bayes rule for each feature, we use the ratio estimates to transform the original feature measurements. Subsequently, we invoke penalized logistic regression, taking as input the newly transformed or augmented features. This procedure trains models with high local complexity and a simple global structure, thereby avoiding the curse of dimensionality while creating a flexible nonlinear decision boundary. The resulting method is called Feature Augmentation via Nonparametrics and Selection (FANS). We motivate FANS by generalizing the Naive Bayes model, writing the log ratios of joint densities as a linear combination of those of marginal densities. It is related to generalized additive models, but has better interpretability and computability. Risk bounds are developed for FANS. In numerical analysis, FANS is compared with competing methods, so as to provide a guideline on its best application domain. Real data analysis demonstrates that FANS performs very competitively on benchmark email spam and gene expression data sets. Moreover, FANS is implemented by an extremely fast algorithm through parallel computing.

Spectral Clustering

Jiashun Jin
Carnegie Mellon University

Consider a two-class clustering problem where we have $(X_i; Y_i), 1 \leq i \leq n,$ from two possible classes. The $X_i$'s are $p \times 1$ vectors that are observable, and $Y_i$ in $\{-1, 1\}$ are class labels which are unknown to us and it is of interest to estimate them.

We propose the following approach to spectral clustering:
1. We use Kolmogorov-Smirnov statistic to assess the importance of the features (i.e., genes).

2. Based on the $p$-values, we perform a feature selection, where the threshold is determined by the recent idea of Higher Criticism Thresholding (HCT). HCT was proposed before for classification, and we must modify it carefully for clustering.

3. Based on all retained features, we obtain the leading eigenvector of the so-called dual covariance matrix, and predict the class labels by the signs of the coordinates of this eigenvector.

We reveal a surprising connection between the HCT and the so-called Signal Noise Ratio (SNR) associated with the post-screening dual empirical covariance matrix. We apply the approach to several gene microarray data sets, where it gives much more satisfactory results than existing clustering methods.

Bayesian Spatial Variable Selection for Ultra-High Dimensional Neuroimaging Data: A Multiresolution Approach

Jian Kang

Emory University

Ultra-high dimensional variable selection has become increasingly important in analysis of neuroimaging data. For example, in the Autism Brain Imaging Data Exchange (ABIDE) study, neuroscientists are interested in identifying important biomarkers for early detection of the autism spectrum disorder (ASD) using high resolution brain images that include hundreds of thousands voxels. However, most existing methods are not feasible for solving this problem due to their extensive computational costs. In this work, we propose a novel multiresolution variable selection procedure under a Bayesian probit regression framework. It recursively uses posterior samples for coarser-scale variable selection to guide the posterior inference on finer-scale variable selection, leading to very efficient Markov chain Monte Carlo (MCMC) algorithms. The proposed algorithms are computationally feasible for ultra-high dimensional data. Also, our model incorporates two levels of structural information into variable selection using Ising priors: the spatial dependence between voxels and the functional connectivity between anatomical brain regions. Applied to the resting state functional magnetic resonance imaging (R-fMRI) data in the ABIDE study, our methods identify voxel-level imaging biomarkers highly predictive of the ASD, which are biologically meaningful and interpretable. Extensive simulations also show that our methods achieve better performance in variable selection compared to existing methods.
Empirical Likelihood in high dimensions

S.N. Lahiri

North Carolina State University

In this talk, we consider some of the challenges associated with extending the Empirical Likelihood (EL) methodology of Owen (1988) to high dimensions. We present results from some recent work on the problem that provide important insights into the technical issues that lead to the failure of the EL. Some potential solutions and their properties will also be presented.

Modeling time-varying effects for high-dimensional covariates: a new Gateaux-differential boosting approach

Yi Li

University of Michigan

Survival models with time-varying effects provide a flexible framework for modeling the effects of covariates on event times. However, the difficulty of model construction increases dramatically as the number of variable grows. Existing constrained optimization and boosting methods suffer from computational complexity. We propose a new Gateaux differential-based boosting procedure for simultaneously selecting and automatically determining the functional form of covariates. The proposed method is flexible in that it extends the gradient boosting to functional differentials in general parameter space. In each boosting learning step of this procedure, only the best-fitting base-learner (and therefore the most informative covariate) is added to the predictor, which consequently encourages sparsity. In addition, the method controls smoothness, which is crucial for improving predictive performance. The performance of the proposed method is examined by simulations and by application to analyze the national kidney transplant data.
Leveraging in Big Data Regression

Ping Ma

Department of Statistics

University of Georgia

Advances in science and technology in the past a few decades have led to big data challenges across a variety of fields. Extraction of useful information and knowledge from big data has become a daunting challenge to both the science community and entire society. To tackle this challenge requires major breakthroughs in efficient computational and statistical approaches to big data analysis.

In this talk, I will present some leveraging algorithms, which make a key contribution to resolving the grand challenge. In these algorithms, by sampling a very small representative sub-dataset using smart algorithms, one can effectively extract relevant information of vast data sets from the small sub-dataset. Such algorithms are scalable to big data. These efforts allow pervasive access to big data analytics especially for those who cannot directly use supercomputers. More importantly, these algorithms enable massive ordinary users to analyze big data using tablet computers.

Classes of Multiple Decision Functions Controlling FWER and FDR

Edsel Pena

University of South Carolina

Two general classes of multiple decision functions, where each member of the first class strongly controls the family-wise error rate (FWER), while each member of the second class strongly controls the false discovery rate (FDR), will be described. These classes offer the possibility that optimal multiple decision functions with to a pre-specified Type II error criterion, such as the missed discovery rate (MDR), could be found which control the FWER or FDR Type I error rates. [This is joint work with Joshua Habiger and Wensong Wu.]
Testing Association without Calling Genotypes Allows for Systematic Differences in Read Depth between Cases and Controls

Yijuan Hu, H. Richard Johnston, Peizhou Liao, Andrew S. Allen, Glen A Satten*

CDC, Atlanta

The quality of genotype calling for next-generation sequence data depends on read depth. Loci with high coverage can typically be called reliably, while those with low coverage may be difficult to call. In an association study, if case participants are sequenced to a greater depth than controls, the difference in genotype quality can introduce a systematic bias. This can easily occur when historical controls (e.g., data from The 1000 Genomes Project) are used as controls. We show how to address this bias by directly comparing the proportion of calls for the minor allele between cases and controls, rather than comparing genotypes. We show that tests based on comparing proportions of calls are valid even in the presence of systematic differences in coverage rate between cases and controls in situations where tests based on genotype have inflated size. We also demonstrate that power gains are possible using designs where we increase the number of controls while decreasing the read depth (while keeping total reads constant).

Safe Methods and Sample Size Determination for Non-Inferiority Trial with Survival Response

Debajyoti Sinha

Florida State University

Recently, there is a substantial interest in the bio-pharmaceutical industry to develop new treatments that may have crucial advantages regarding administration, cost, and tolerability compared to a current treatment of proven efficacy. An essential step for approving such a new treatment is to conduct a non-inferiority trial to evaluate whether the new treatment is at least as
efficacious as the existing treatment in terms of the main survival outcome. Here, we first investigate the serious consequences of usual log-rank based non-inferiority tests when the proportional hazard model (PHM) assumption fails. Then we introduce a practical formulation of the non-inferiority hypothesis as well as the corresponding statistical procedures, both frequentist and Bayesian, based on proportional odds survival model (POSM). Unlike the commonly used tests, our procedures ensures a proper control of the type I error rate when two treatment arms satisfy either the POSM or PHM. We also present a practical method for determining sample sizes to guarantee a desired type-II error rate.

Regression analysis of networked data

_Peter Song_

_University of Michigan_

We develop a new regression analysis approach to evaluating associations of covariates with outcomes measured from networks. This development is motivated from a study of infant growth that collects outcomes of event related potentials (ERP, a type of neuroimaging) measured over electroencephalogram (EEG) electrodes on the scalp. We propose a new generalized method of moments (GMM) that incorporates both established and data-driven knowledge of network topology among nodes in the estimation and inference to achieve robustness and efficiency. The GMM approach is computationally fast and stable to handle the regression analysis of network data, and conceptually it is simple with desirable properties in both estimation and inference. Both simulation studies and real EEG data analysis will be presented for illustration.
Determination of sample size for a multi-class classifier based on single nucleotide polymorphisms: A Volume Under the Surface approach

T. N. Sriram
University of Georgia

For coded single-nucleotide polymorphism (SNP) data from \( D(\geq 2) \) classes, we derive an optimal Bayes classifier and a linear classifier, and obtain a normal approximation to the probability of correct classification for each classifier. These approximations are then used to evaluate the associated Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) or Volume Under the ROC hyper-Surface (VUS), whose accuracies are validated using Monte Carlo simulations. We give a sample size determination method, which ensures that the difference between the two approximate AUCs (or VUSs) is below a pre-specified threshold. The performance of our sample size determination method is then illustrated via simulations. For the HapMap data with three and four populations, a linear classifier is built using 92 independent SNPs and the required total sample sizes are determined for a continuum of threshold values. In all, four different sample size determination studies are conducted with the HapMap data, covering cases involving well-separated populations to poorly-separated ones.

Simple estimation procedures for regression analysis of interval-censored failure time data under the proportional hazards model

(Tony) Jianguo Sun
University of Missouri

Interval-censored failure time data occur in many fields including epidemiological and medical studies as well as financial and sociological studies, and many procedures for their analysis have been proposed. In particular, a number of procedures have been developed for regression
analysis of interval-censored data arising from the proportional hazards model (Finkelstein, 1986; Huang, 1996; Pan, 2000). For most of these procedures, however, one drawback is that they involve estimation of both regression parameters and baseline cumulative hazard function. In this talk, we present two simple estimation approaches that do not need estimation of the baseline cumulative hazard function. The asymptotic properties of the resulting estimates are given, and an extensive simulation study suggests that they work well for practical situations.

Cost and Cost-Effectiveness Analysis with Incomplete Data

Hongkun Wang
Georgetown University

Cost and cost-effectiveness study plays an important role in evaluating the economic impact of new treatments, evaluating factors associated with cost, and finding more cost-effective therapies. In long-term clinical or observational studies, patients are recruited over time and the study terminates before all the patients reach the endpoints of interest so that their medical costs are not fully observed. Censoring can also occur due to staggered entry or loss to follow-up. Another difficulty in analyzing cost data is the correlation between censored costs and potential uncensored costs. In this talk we will review and discuss some available methods for estimating the mean cost with censored data, modeling the mean cost, and the cost-effectiveness analysis using iCER.
Sufficient Dimension Reduction in Binary Classification

Yichao Wu

North Carolina State University

Reducing dimensionality of data is essential for binary classification with high-dimensional covariates. In the context of sufficient dimension reduction (SDR), most, if not all, existing SDR methods suffer in binary classification. In this talk, we target directly at the SDR for binary classification and propose a new method based on support vector machines. The new method is supported by both numerical evidence and theoretical justification.

Methods for Differential Methylation Analysis with Next Generation Sequencing

Hongyan Xu and Varghese George

Department of Biostatistics & Epidemiology

Georgia Regents University

A major take-home message from the large number of genome-wide association studies in recent years is that common genetic variants in the primary DNA sequence explains only a small portion of the genetic predisposition to complex diseases. More recently, epigenetic changes, especially DNA methylation at CpG loci, are becoming of interest in dissecting complex diseases since methylation patterns may be representative of environmental exposures important for the disease. With the development of next-generation sequencing, it is now feasible to generate large data to analyze differential methylation for epigenomewide association studies. In this presentation, we describe some of the existing methods for differential methylation at single CpG sites and for CpG regions, and introduce some new approaches being developed by our team. We illustrate our methods using simulated as well as real data.
Nonparametric statistics methods for medical diagnostic marks with right censored data

Yueheng An
Georgia State University

In order to justify the effect of a new medicine or a new cure under the support of statistical output, physicians and medical researchers impose significant concentrations on the comparison of two treatments in clinical trials and related medical studies. The receiver operating characteristic (ROC) analysis has been successfully applied to classifier problems with two classes. The area under the ROC curve (AUC) has been elected as a better way to evaluate classifiers than predictive accuracy or error and has also recently been used for evaluating probability estimators. On the other hand, censoring often occurs when the value of an observation or measurement is outside the range of a measuring instrument or only partially known in statistical and medical studies. Therefore, we are interested in the ROC curves, the difference of two ROC curves, the AUCs and the difference of two AUCs, all with right censored data. In this poster, empirical likelihood method is employed to analysis the performance of the difference of two ROC curves and the difference of two AUCs. Confidence intervals and coverage probabilities show that empirical likelihood has a better performance than normal approximation. Joint work with Yichuan Zhao

Stochastic Simulations of Bio---chemical Reaction Networks

Anilkumar Devarapu
Albany State University

Understanding the chemistry of cells is important to our understanding of living things, from bacteria to humans. In recent years, stochastic simulation has played an increasingly important role in furthering our understanding of biochemical reactions. In this article, we explain both the exact and the approximation stochastic algorithms for analyzing dynamics of biochemical reaction networks.
Signed Rank Inference via the Empirical Likelihood

Frazier Bindele and Yichuan Zhao

University of South Alabama

In this paper, we consider a general regression model and we study the signed-rank estimator of the regression coefficients. Under independent and possibly dependent model errors structure, an empirical likelihood based on the signed rank (ELSR) estimating equation is defined to determine coverage probabilities and construct confidence intervals of the regression coefficients. Monte Carlo simulation experiments are used to compare the normal approximation (NA) and the ELSR methods via coverage probabilities, and show that the proposed ELSR method outperforms the NA method for different settings of the model error.

Signal Estimation Under Random Compositional and Additive Noises

Jason Cleveland and Wei Wu

Florida State University

In statistical signal processing, observations are often assumed as realizations from a parametric or semi-parametric generative models with some ground-truth template. The estimation of such template has been a fundamentally important problem. In particular, this problem is of significant challenge when both compositional and additive noises are present in the model, where the compositional noise indicates time warping along the x-axis, and additive noise indicates conventional error in the y-axis. Estimation when only additive noise is present is a classical problem and can be easily solved using standard large-sample theory. A recent study that extends the well-known Fisher-Rao metric demonstrates that when time warping is present, the underlying template can also be effectively estimated from a set of smooth observations. In this paper, we extend this work by building a generative model for the random observations where both noises are present. We provide an algorithm to compute the underlying template, and mathematically demonstrate that the estimator is consistent asymptotically. As compared to the extended Fisher-Rao method, the new framework is immune to the additive noise and is therefore free of the smoothness assumption. We also find that the new framework can better identify underlying structure and provide superior classification performance in simulations as well as a real SONAR dataset.
A Score-type Test for Heterogeneity in Zero-inflated Models in a Stratified Population

Guanqun Cao
Auburn University

We propose a score-type statistic to evaluate heterogeneity in zero-inflated models for count data in a stratified population, where heterogeneity is defined as instances in which the zero counts are generated from two sources. In this paper, we extend the literature by describing a score-type test to evaluate homogeneity against general alternatives that do not neglect the stratification information under the alternative hypothesis. Our numerical simulation studies show that the proposed test can greatly improve efficiency over tests of heterogeneity that ignore the stratification information. An empirical application to dental caries data in early childhood further shows the importance and practical utility of the methodology in using the stratification profile to detect heterogeneity in the population.

High-Dimensional Canonical Forest

Yu-Chuan Chen
National Center for Toxicological Research, FDA

Recently a new ensemble classification method named Canonical Forest has been proposed by Chen et al. (2014). Although it has been proven to be a comparable classification ensemble method by giving consistently good results in many data sets in their experiment, it requires adopting feature selection method before applying to high-dimensional data. Here we present a variant version of Canonical Forest called High-Dimensional Canonical Forest (HDCF) that can directly work on high-dimensional data without any features pre-selection by implementing Random Subspace method into the algorithm of Canonical Forest. We conducted an experiment using data sets on gene imprinting, estrogen and leukemia to compare HDCF with Random Forest, CERP, and SVM which all are popular and successful classification methods on high-dimensional data sets. In our experiments, HDCF had the highest classification accuracy in gene imprinting and leukemia data sets and second highest classification accuracy in estrogen data set. Besides the classification accuracy, we also investigated the balance between sensitivity and specificity for all these four classification methods. The performance of HDCF on the balance between sensitivity and specificity was quite comparable to the other three classification methods.
A Bayesian approach to select multi-level pharmacodynamic markers of kinase inhibitor using phosphoproteomics data

Yian (Ann) Chen
Moffitt Cancer Center

Dysregulation of signaling networks plays a critical role in cancer biology. Mass-spectrometry based phosphoproteomics profiling has enabled the identification and quantification of proteins in the signaling network. We have developed a Bayesian approach to select subsets of markers at multiple levels (e.g., phosphopeptides, proteins, or pathways) using latent variables and evaluate the joint effects of selected combinations via stochastic search for predicting drug efficacy. We illustrated this in two settings, one to integrate experimental phosphopeptide quantifications with protein information and the other one to integrate experimental phosphopeptide quantifications with biological pathway and phospho-site-specific functional information. The quantitative phosphoproteomics data were generated using anti-phosphotyrosine peptide affinity purification followed by liquid chromatography-mass spectrometry (LC-MS/MS) after treating a human sarcoma cell line A204 with one of the kinase inhibitors, dasatinib and imatinib. Ranked by posterior probabilities, some of the highly ranked proteins including FYN, YES, EPHA2, PGFRA, and GAB1, are known to be drug targets; while the roles for other highly ranked proteins, such as VIME and P85B (PI3K regulatory subunit beta), are less clear. By integrating the phosphoproteomics quantification with the pathway information, we identified additional proteins and phosphotyrosine sites, such as MK01 and MK03, whose phosphorylation levels were associated with drug efficacy, in an unexpected direction. We have validated their relative phosphorylation levels using quantitative western blotting. Phosphorylation patterns were different for some of the pYs within the same protein supporting the need for a multi-level analysis. Our approach selected markers jointly across levels for predicting kinase inhibitor efficacy in a unified framework.
Principal component analysis (PCA) is a well-established technique for quantifying systematic variation in high-dimensional data with $m$ variables and $n$ observations. In modern applications of PCA in genomics and population genetics where $m \gg n$, we often consider principal components (PCs) as non-parametrically estimating latent variables. PCs are computed as weighted sums of $m$ variables, and this set of weights – called the sample loadings – are constructed in order to maximize sample variance. While reliable PCs are constructed by adding up small contributions from $m$ variables, each of the sample loadings is nonetheless imprecise. We have developed a significance-based shrinkage approach to obtain accurate and sparse loadings of PCs. Non-parametrically estimated $p$-values for association between $m$ variables and PCs are used either (a) to estimate a proportion of null variables and hard-threshold the same proportion of the sample loadings or (b) to estimate $m$ posterior inclusion probabilities to downweight the sample loadings accordingly. Using a wide range of simulation scenarios, we demonstrate that the proposed shrinkage methods substantially improve the accuracy of loadings, the proportion variance explained, and the covariance matrix. As an application, we considered the consumption and release profiles of metabolites in NCI-60 cancer cell lines (Jain et al., 2012). The proposed shrunken covariance matrix was used to construct a metabolite association network, which provides an insight into metabolic reprogramming activities inherent to cancer. Overall, in the context of high-dimensional latent variable models, this paper introduces a coherent set of shrinkage methods to better estimate PCA and covariance matrix.
New modality of breast cancer prognosis – using complex wavelet on mammogram image classification

Seonghye Jeon

IBM Global Business Services/ Advanced Analytics and Optimization

Breast cancer is one of the most common forms of cancer among women in the United States; an estimated 1 in 8 women born today will be diagnosed during her lifetime. Since the causes of breast cancer are not yet fully understood, early detection is still the best strategy for improving prognoses. Mammography is currently the most effective method for detecting breast cancer early; however, radiological interpretation of mammogram images is a challenging task. The appearance of even normal tissue is highly variable and complex, and signs of early disease are often indistinct.

We propose a diagnostic based on the properties of overall image backgrounds; this procedure currently is an unused diagnostic modality in mammograms. Generally speaking, normal/healthy images tend to be more irregular than cancer images. The overall regularity of the image is assessed through wavelet analysis, which is then summarized by a few measures. These estimators are evaluated based on their ability to classify digitized mammogram images from a clinical database, for which the true disease status is known by biopsy.

This research presents a two-fold approach: (i) generalization of the covariance wavelet spectra to the complex domain and (ii) the estimation of Hurst parameter and phase information as discriminatory descriptors. The most accurate classification rates from this work achieve 85%; these rates vary slightly with the choice of wavelet basis, levels used and size of training set.
Kulkarni and Rattihalli (2002) proposed an estimator for a bivariate mean residual life function that is shown to be uniformly strongly consistent, and converges weakly to a zero-mean Gaussian process on proper normalization. In this article, we apply Empirical Likelihood (EL) method to the two-dimension estimator and perform simulation studies with the same settings such as sample sizes and methods used to generate random sample mentioned in Kulkarni and Rattihalli, to evaluate the coverage probabilities and do comparison study with Normal Approximation (NA). Coverage probability under EL is found to be closer to nominal levels and better than NA. In addition, we used profiling method to study the estimator in one dimension, applied EL and performed simulation studies with the same settings as mentioned in Kulkarni and Rattihalli (2002). EL was found closer to the nominal levels and better than NA. Joint work with Yichuan Zhao.

SVSI: A fast and powerful set-valued system identification approach to identifying rare genetic variants for ordered categorical phenotype

Wenjian Bi¹*, Guolian Kang²**, Yanlong Zhao¹, Yuehua Cui³, Yun Li⁴, Christine M Hartford⁵, Wing Leung⁵, Ji-Feng Zhang¹

¹Key Laboratory of Systems and Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, P.R.C.
²Department of Biostatistics, St. Jude Children’s Research Hospital, Memphis, Tennessee 38105, U.S.A.
³Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, U.S.A.
For phenotype-genotype association studies that involve a phenotype with ordered multiple response categories, we usually either regroup multiple categories of the phenotype into two categories of “cases” and “controls” and then apply the standard logistic regression (LG) model [1-2], or apply non-parametric method of spearman rank correlation [3] or parametric method of ordered logistic (orderLG) regression model [4] which accounts for the ordinal nature of the phenotype. However, these approaches may lose statistical power if the phenotype is obtained by categorizing an observed or complicated unmeasured or immeasurable continuous phenotype or if the underlying genetic variants are rare. Therefore, we propose a set-value (SV) system method, which assumes that the underlying continuous phenotype follows a normal distribution, to identify genetic variants associated with an ordinal categorical phenotype. We couple this model with a set-value system identification method to identify all underlying key system parameters. Simulation studies show that SV well controlled the Type I error rate. In the comparison among LG, SV and orderLG methods, LG had significantly lower power than both SV and orderLG due to the disregard of the ordinal nature of the phenotype, and SV had similar or higher power than orderLG. Additionally, the SV association parameter estimate was 2.7-28.7 fold less variable than the orderLG association parameter estimate. Less variability in the association parameter estimate translates to greater power and robustness across the spectrum of minor allele frequencies. These advantages are most pronounced for rare variants or even common variants when sample size is small. For instance, in a simulation with data generated from an additive orderedLG model with odds ratio of 7.4 for a phenotype with three categories, a single nucleotide polymorphism with minor allele frequency of 0.75% and sample size of 999 (333 per category), the power of SV, orderLG and LG models were 70%, 40% and <1%, respectively, at a significance level of 10^-6. When applied to a real data set, the set of variants identified by LG and orderLG was a subset of those identified by SV. Thus, SV can be a competitive alternative to LG or orderLG in genetic association studies such as candidate gene, genome-wide association studies or next generation sequencing studies, for ordered categorical phenotype.

References


---

**Non-decimated 2-D wavelet spectra and its use in Breast Cancer Diagnostics**

*Minkyoung Kang, Brani Vidakovic*

*Georgia Institute of Technology*

Breast cancer is the second leading cause of death from cancer among women in USA. Mammography diagnosis has been the only way for the early detection of it, but it often misses cancer detection due to indistinctive symptoms to human eyes. Instead of human screening, we used self-similarity property of medical images for classification. 2D non-decimated wavelet transformation produces a comparatively accurate self-similarity index, as it generates redundant wavelet coefficients at each decomposition level providing more samples in averaging. The mammography diagnosis via 2D non-decimated wavelet transformation achieves 65 to 70% of classification accuracy.
Modeling Complex Gating Properties of the Human Sodium Channel

Konstantin G. Kapustin and Vladimir E. Bondarenko

1Department of Mathematics and Statistics and 2Neuroscience Institute, Georgia State University

Sodium channel is the major channel that causes depolarizations of excitable cells, such as neurons and cardiac myocytes. It is vital to model the gating properties of this channel in order to predict electrophysiological behavior of the wild type and mutant channels. The present study introduces several Markov models of the Na+ channel with various (open, closed, and inactivated) states and coupling between them. Simulated voltage-clamp protocols were used to determine the activation, deactivation, steady-state inactivation, and recovery from inactivation kinetics, as well as current-voltage relationships, steady-state inactivation, and voltage dependence of normalized channel conductance. The study demonstrates the necessity of coupling the fast and the slow inactivated states in order to reproduce the complex gating properties of the human Na+ channel which were found experimentally.

Symbolic Data Analysis of Progression on Hepato Cellular Carcinoma

Tae Rim Lee

Dept. of Bioinformatics & Statistics,
Korea National Open University

Viral Hepatitis B infection is one of the major infectious diseases with more than 360 million chronic carriers worldwide and Korea is an endemic area of chronic hepatitis B virus (HBV) infection. As one of efforts to search for host genetic factors influencing the outcome of chronic HBV infection, especially the progression to HCC, SNPs of several candidate genes were discovered and screened.
The purpose of this study was designed to evaluate the prognosis of HCC in relation to treatment method and their affecting gene and clinical factors by symbolic data analysis. Proposed SDA model shows the genes related risk factors of HCC. These findings could be available to predict prognosis of HCC and give valuable information to justify the treatment strategy for clinician. The Symbolic Data Analysis approach aim to study classes of individuals considered as new units. SDA actually provides a new perspective on this study and suggests that some genetic and clinical parameters are related to the HCC and its progression. Finally, the results obtained are compared to those previously provided by tree structured survival model.

**Generalized confidence intervals for partial youden index**

**and its corresponding optimal cut-off point**

*Chenxue Li*

*Georgia State University*

In the field of diagnostic test studies, the accuracy of a diagnostic test is essential in evaluating the performance of the test. The receiver operating characteristic (ROC) curve and the area under the curve (AUC) are widely used in such evaluation procedures. Meanwhile, the Youden index is also introduced into practice to measure the accuracy of the diagnostic test from another aspect. The Youden index maximizes the sum of sensitivity and specificity, assuring decent true positive and negative rates. It draws one's attention due to its merit of finding the optimal cut-off points of biomarkers. Similar to Partial ROC, a new index, called “Partial Youden index” can be defined as an extension of Youden's Index. It is more meaningful than regular Youden index since the regular one is just a special case of the Partial Youden Index. In this thesis, we focus on the construction of generalized confidence intervals for the Partial Youden Index and its corresponding optimal cut-off points. Extensive simulation studies are conducted to evaluate the finite sample performances of the new intervals.
Minimum distance model checking when responses are missing at random

Xiaoyu Li
Auburn University

This paper proposes a class of lack-of-fit tests for fitting a linear regression model when some response variables are missing at random. These tests are based on a class of minimum integrated square distances between a kernel type estimator of a regression function and the parametric regression function being fitted. These tests are shown to be consistent against a large class of fixed alternatives. The corresponding test statistics are shown to have asymptotic normal distributions under null hypothesis. Some simulation results are also presented.

Influence Function-based Empirical Likelihood Inferences for Lorenz Curve

Bing Liu
Georgia State University

In this thesis, an empirical likelihood method based on influence function is developed and used to construct confidence intervals for the Lorenz ordinates. This method is defined under the simple random sampling and the limiting distribution of the proposed empirical likelihood ratio statistic is a standard Chi-square distribution. Extensive simulation studies are conducted to evaluate the proposed empirical likelihood-based confidence intervals for the Lorenz ordinates. Finally, this method is used on a real income data as an application.
Statistical Inference for Frailty-Induced Correlated Competing Risks Data

Piaomu Liu

Department of Statistics, University of South Carolina-Columbia

In this paper, distributional properties and some surprising results pertaining to recurrent event models based on marked Poisson processes are obtained. In this competing risks model scenario, event occurrences are assumed to follow homogeneous Poisson processes (HPPs), both without and with frailties. Each risk is observed over a random window. Properties of the number of events seen over the monitoring time window, and the gap-time that covers the termination time are examined for each risk. Correlation among gap times of the observed events and the gap times covering the monitoring time across different risks are also derived. Seemingly peculiar-looking results induced by the sum-quota accrual scheme are highlighted and discussed. An EM algorithm is developed for parameter estimation. Through the simple marked Poisson process with recurrent events, it is hoped that a better appreciation for difficulties in dealing with recurrent events will rise.

Jackknife Empirical Likelihood for Estimating Variance without Estimating Mean Function

Yuan Liu and Yichuan Zhao

Georgia State University

The variance of a random variable is $\sigma^2$. It is the measure of spread from the center. Therefore, how to accurately estimate $\sigma^2$ has always been an important topic in recent years. How to
accurately estimate $\sigma^2$ has always been an important topic in recent years. In this paper, we used the jackknife empirical likelihood (JEL) method to estimate variances without estimation of the mean function in nonparametric regression models. The log-ratio of the proposed JEL converges to the standard chi-squared distribution. The simulation study is carried out to compare the JEL method and standard methods in terms of coverage probability and interval length for the confidence intervals from different models. The proposed JEL has better performance. She also illustrates the proposed methods using two real data sets.

**Performance of the Jackknife Empirical Likelihood and Inference for the Concordance Correlation Coefficient**

*Anna Susan Moss*

*Georgia State University*

Lin's Concordance Correlation Coefficient ($\rho_c$) is a commonly used measure of reproducibility or agreement in paired samples. Interval estimates for ($\rho_c$) are typically constructed using the normal approximation method (NA) proposed by Lin (1989). In this work, we evaluated three empirical likelihood methods, Jackknife Empirical Likelihood (JEL), adjusted JEL (AJEL), and extended JEL (EJEL), for interval estimation. We compared performance of the 3 JEL methods to the standard NA method by computing coverage probability and average interval length from paired samples generated from normal and non-normal distributions (exponential, Poisson, and uniform). We also provide an application of the JEL methods using actual data.
Heterogeneity of contraction in mouse ventricular myocytes: Modeling study

Paula D. Mullins and Vladimir E. Bondarenko

Department of Mathematics and Statistics and Neuroscience Institute, Georgia State University

Cells within a heart muscle are not homogeneous, they vary based on several factors. One of these factors is their location. The model presented is for cells in the left ventricular region, but experiments have shown that even in a specific region there are differences between epicardial and endocardial cells. In particular, endocardial cells have a more prolonged action potential and a larger calcium transient than epicardial cells. We have modified our previous model for mouse ventricular myocyte contraction to include the differences in action potential and calcium transients which have been found in epicardial versus endocardial cells. The model was developed for room temperature (+25 C) and consists of 51 differential equations. The model was implemented in Fortran and the differential equations were solved by the 4th-order Runge-Kutta method with the time step 0.0001 ms. Our simulations fit well with previously published experimental data for the heterogeneity in action potential duration, calcium transients, and potassium currents. We were also able to match the force development and cell shortening of mouse ventricular cells. While there is no experimental data on heterogeneity of contraction in mouse ventricle, our model predicts the heterogeneous contraction force based on the heterogeneity of calcium transients.

Semiparametric Bayesian Inference of Gap-Time Distribution with Recurrent Event Data

A.K.M. Fazlur Rahman

Department of Statistics, University of South Carolina, Columbia

Recurrent event data arise from a wide variety of studies/fields such as clinical trials, epidemiology, public health, and biomedicine (e.g. repeated heart attack, repeated tumor occurrences of a cancer patient). Semiparametric Bayes inference of the gap-time survivor
function with the effect of covariates of a correlated recurrent event in the presence of censoring is considered. A frailty model is considered to allow the association between inter-occurrence gap-times. We assume that for a subject or unit given the unobserved frailty variable \( W = w \), the inter-occurrence gap-time \( \{T_j\} \) are IID with some distribution function \( F(t \mid W = w) \). In our procedure, we assign a Gamma process prior on the baseline cumulative hazard function \( \Lambda_0 \) and parametric prior distributions on the finite dimensional parameters associated with covariates and frailty. We derive the conditional posterior distributions from the joint posterior distribution of the unknown parameters of interest and employ the Gibbs sampler techniques to obtain samples from the joint posterior distribution. Simulation studies demonstrate the effectiveness of the developed method. Peña et al. (2001) estimator of \( \bar{F} \) for a correlated recurrent event data without covariates is a special case of our developed estimator with the precision parameter of the Gamma process prior tends to zero.

Joint work with Dissertation advisor Dr. Edsel Peña.

**A Modified Warner’s Randomized Response Model**

*Jeong Sep Sihm*

*The University of North Carolina at Greensboro*

We propose to generalize the usual Warner’s RRT model by introducing a modified Two-Stage Binary Optional Randomized Response Technique (RRT) model which gives respondents the freedom to choose whether to provide a scrambled response or a true response. Similar model was proposed by Sihm and Gupta (2014) in the context of Unrelated Question RRT Models but using a split sample approach. In this poster, an alternative to the split sample method is used to estimate the prevalence of the sensitive characteristic as well as sensitivity level of the underlying sensitive question. The proposed model is compared with the model that uses the split sample approach, and the original Warner’s model. Computer simulations show that the new model has the smaller variances for the two parameters when the sample size is equal and can give reliable estimates of \( W \) when the split sample approach breaks down in case of \( \pi = 0.5 \).
The Pietra ratio (or Robin Hood Index) is a measure of inequality in a resource distribution in population, it is a ratio of mean deviation to 2 times mean (\( P = \frac{\sum_{i=1}^{n} |x_i - \mu|}{2\mu} \)). In this paper, we develop Bootstrap methods, Jackknife Empirical Likelihood (JEL; 2009, Jing et al), and two extended methods of JEL-Adjust Empirical Likelihood (AEL; 2008, Chen et al) and Extend Empirical Likelihood (EEL; 2013, Tsao) for estimating Pietra ratio in different distributions. We compare these methods in terms of the coverage probabilities and interval lengths of Pietra ratio, and we found that AEL and EEL methods perform better than the other methods. Several real data sets are used to check the conclusions which come from the simulation study.

Fitting Flexible Parametric Regression Models with GLDreg in R

Steve Su
Covance Pty. Ltd.
Macquarie University Research Park

This presentation outlines the functionality of GLDreg package in R which fits flexible parametric regression models using generalised lambda distributions (GLDs) via maximum likelihood estimation and L moment matching. Once the GLD regression model is obtained, parametric quantile regression model can then be estimated. The quality of regression model is assessed using QQ plots and Kolmogrov-Smirnoff goodness of fit test and an overall summary plot of regression coefficients are also given as part of the regression model output. The main
advantage of GLDreg is the provision of robust regression lines and smooth regression quantiles beyond the capabilities of existing known methods. This work is a direct implementation of Su (In Press) and uses GLDEX 2.0.0.0 (Su 2007, Su 2010).

References


Improving Trauma Triage Models for Motor Vehicle Crashes

Yaoyuan Vincent Tan

University of Michigan, School of Public Health, Biostatistics

Severe injury prediction in motor vehicular crashes has long been an interest of trauma researchers. Many studies have shown that delta-v, the vehicle velocity associated with the primary direction of force of the crash event, is a strong predictor for severe injury outcomes. Delta-v is often estimated by re-constructing the accident after crash investigations. However, this process has been shown to underestimate delta-v. Recent advances in technology have enabled a device, Event Data Recorders (EDR), to capture the full deceleration trajectory during crashes when air-bags are deployed. Our method proposes to use a recently developed statistical technique called functional data analysis (FDA) in order to fully utilize this deceleration profile. FDA captures information in trajectories using functions and then applies standard statistical techniques to them. Our method is different in two aspects: first, we propose the use of the slope of the deceleration profile, a variable highly correlated with delta-v, in the injury prediction model, and second, we advocate retaining the information not captured by the functions during FDA. We show in our results that retaining this information enhances our prediction model. We
then apply our method to 2005-2011 EDR data sets available on the National Highway and Transportation Safety Administration (NHTSA) website. We enhance our prediction model with baseline covariates found in National Automotive sampling System (NASS) Crashworthiness Data System (CDS). Our model performed well but could have benefited from a larger sample size. Our results will be useful for Emergency Medical Services (EMS) looking to improve their response to motor vehicular crashes and will ultimately enhance severe injury prediction.

**Extended Matrix and Inverse Matrix Methods Utilizing Internal Validation**

**Data When Both Disease and Exposure Status Are Misclassified**

*Li Tang*

*St. Jude Children’s Research Hospital*

The problem of misclassification is common in epidemiological and clinical research. In some cases, misclassification may be incurred when measuring both exposure and outcome variables. It is well known that validity of analytic results (e.g., point and confidence interval estimates for odds ratios of interest) can be forfeited when no correction effort is made. Therefore, valid and accessible methods with which to deal with these issues remain 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 well-studied “matrix” and “inverse matrix” methods into the framework of maximum likelihood, our approach allows the flexible modeling of a richer set of misclassification mechanisms when adequate internal validation data are available. The value of our extensions and a strong case for the internal validation design are demonstrated by means of simulations and analysis of bacterial vaginosis and trichomoniasis data from the HIV Epidemiology Research Study (HERS).
MULTIPLE INFLATION (MI) COUNT MODEL S: BEYOND ZERO INFLATED (ZI) MODELS

Arvind Tripathi
University of Alabama at Birmingham

Now-a-days, Zero Inflated (ZI) models are fairly common in use. The Google scholar search returns 4240 articles for years 2012-2013 on “Zero Inflated”, which underscores the importance of inflated count models when only one count zero is inflated. However, difficulties arise when the outcome variable has more than one inflated counts. Till date, no model provides flexibility to model such data precisely. We proposed Multiple Inflation (MI) count regression model by using mixture of cumulative logit and Poisson/Negative Binomial model to analyze count outcome variable with the multiple inflation. An EM algorithm is used to obtain the maximum-likelihood estimates and the proposed model is compared with the other available competitive count models.

Lag selection for single-index time series models

Guannan Wang
University of Georgia

Many of the time series data exhibit nonlinearity. Nonlinear autoregressive models are very popular to use in time series analysis because of its great flexibility. In such modeling process, very often one needs to include many lagged variables in the model to capture the persistence of a time series. Sometimes, the lag length can be very long, or even close to the length of time series. Such “curse of dimensionality" problem is challenging for nonparametric modeling and there is a need to select significant explanatory lagged variables. Single-index model is an appealing and fundamental tool for handling “curse of dimensionality". In this paper we consider nonlinear single-index time series models. In addition we propose a method to select significant explanatory lagged variables. We apply polynomial spline basis function expansion and smoothly clipped absolute deviation penalty to perform estimation and lag selection in the
framework of high-dimensional time series. Under stationary and strong mixing conditions, the resulting estimators enjoy the “oracle” property even when the number of index parameters tends to infinity as the sample size increases. An efficient iterative algorithm has been developed to identify the lags and estimate the coefficients simultaneously. Both numerical studies and real data application confirm a good performance of the proposed method.

**Oxford Parkinson’s Disease Detection data set: Supervised and Unsupervised Learning**

*Xiaoyuan Wang*

*Georgia State University*

Biological research is becoming increasingly database drive and statistical learning can be used to discover trends and patterns in the underlying data and interpret the bio-information. Both supervised and unsupervised learning approaches can often be used to analyze the same kind of data for different purposes. The Oxford Parkinson’s disease detection data is employed in this project for the purpose of examining multiple data mining methodologies. My objective is to find relationships that predict the Parkinson’s disease detection and discover pattern by organizing data into clusters.

INDEX WORDS: Supervised Learning, Unsupervised learning, Discriminate Analysis, SVM, Decision Trees, KNN, K mean, Hierarchical Clustering, Mixture models.

**Characterizing Exons and Introns by Regularity of Nucleotide Strings**

*Tonya Woods*

*Georgia Institute of Technology*

In this work, we outline a methodology for transcribing sequences of DNA to numeric matrices in order to analytically investigate important structural characteristics of DNA. This methodology involves assigning unit vectors to nucleotides, placing the vectors into columns of a matrix, and accumulating across the rows of this matrix. Transcribing the DNA in this way allows us to compute the 2-D wavelet transformation and assess regularity characteristics of the sequence via the slope of the wavelet spectra. Based on our discovery of equivalence classes, we
adjust the methodology so that the log spectral slope does not depend on the assignment of unit vectors. In addition to computing a global slope measure for a sequence, we can apply our methodology for overlapping sections of nucleotides to obtain an evolutionary slope. To illustrate our methodology, we analyze 376 gene sequences from the first chromosome of the honey bee. First, we simulate “random DNA” using actual proportions of the nucleotides from the honey bee DNA and find that the actual DNA sequence is more regular than simulated DNA sequences. The second analysis involves calculation of the cumulative evolutionary slope for each gene. We label each window of DNA as exons (coding regions), introns (noncoding regions), or a combination of the two and average the cumulative evolutionary slopes for each of the three categories. For the genes analyzed, we find that introns are significantly more regular (lead to more negative slopes) than exons, which agrees with the results from the literature where regularity is measured on “DNA walks.”

Generalized Smoothed Quantile-based Likelihood Ratio Tests on Nonparametric Density Alternatives

Han Yu

Northwest Missouri State University

A new kernel-type of quantile-based likelihood ratio tests is proposed for nonparametric density alternatives. The proposed tests are constructed to be asymptotically distribution-free for the test of a parametric null hypothesis against the nonparametric density alternatives. The limiting distribution that quantifies the convergence rate is derived. The procedure can be viewed as a novel nonparametric extension of the classical parametric likelihood ratio test. Simulations for the finite sample size performance of the proposed tests are provided.
**FMEM: functional mixed effects modeling with applications to the analysis of longitudinal white matter tract statistics**

*Ying Yuan*

*St. Jude Children’s Research Hospital*

Motivated by recent work studying massive neuroimaging data from longitudinal studies, we propose a functional mixed effects model (FMEM) for modeling varying association function between repeated functional responses and a set of covariates of interest, while accounting for complex spatial-temporal correlations. FMEM consists of a functional mixed effects model, an efficient method for spatially smoothing varying coefficient functions, an estimation method for estimating the spatial–temporal correlation structure, a test procedure with local and global test statistics for testing hypotheses of interest associated with functional response, and a simultaneous confidence band for quantifying the uncertainty in the estimated coefficient functions. Simulated data are used to evaluate the finite sample performance of FMEM and to demonstrate that FMEM significantly outperforms the standard point-wise mixed effects modeling approach. We apply FMEM to study the spatial–temporal dynamics of white-matter fiber tracts in a clinical diffusion tensor imaging study of neurodevelopment.

**A Non-parametric Model to Address Overdispersed Longitudinal Count Responses with Missingness**

*Hui Zhang*

*St. Jude Children’s Research Hospital*

Although widely used for comparing multiple samples in biomedical research, the analysis of variance (ANOVA) model suffers from a series of flaws that not only raise questions about conclusions drawn from its use, but also undercut its many potential applications to modern clinical and observational research. In this poster, we propose a new class of generalized ANOVA models to concurrently address all these fundamental flaws underlying this popular multi-group comparison approach so that it can be applied to many immediate as well as
potential applications ranging from addressing an age-old technical issue in applying ANOVA to cutting-edge methodological challenges arising from the emerging effectiveness research paradigm. By integrating the classic theory of U-statistics with state-of-the-art concepts such as the inverse probability weighted estimates, we develop distribution-free inference for this new class of models to address missing data for longitudinal clinical trials and biomedical studies, especially for Next Generation Sequencing (NGS) data. We illustrate the proposed class of models with both real and simulated study data, with the latter investigating behaviors of model estimates under small and moderate sample sizes.

**Jackknife Empirical Likelihood Inference for the Skewness and Kurtosis**

*Yan Zhang and Yichuan Zhao*

*Department of Mathematics and Statistics*

*Georgia State University*

Skewness and kurtosis are measures used to describe shape characteristics of distributions. In this thesis, we examine the interval estimates about the skewness and kurtosis by using jackknife empirical likelihood (JEL), adjusted JEL, extended JEL, traditional bootstrap, percentile bootstrap, and BCa bootstrap methods. The limiting distribution of the JEL ratio is the standard chi-squared distribution. The simulation study of this thesis makes a comparison of different methods in terms of the coverage probabilities and interval lengths under the standard normal distribution and exponential distribution. The proposed adjusted JEL and extended JEL perform better than the other methods. Finally we illustrate the proposed JEL methods and different bootstrap methods with three real data sets.

INDEXWORDS: Skewness, Kurtosis, Empirical likelihood, Jackknife empirical likelihood, Adjusted jackknife empirical likelihood, Extended jackknife empirical likelihood, Bootstrap, Bootstrap percentile, Bootstrap BCa, Coverage probability, Interval length
A Bioinformatics pipeline for next generation sequencing data analysis of *Salmonella* foodborne pathogen

*Weizhong Zhao*¹,², *James J. Chen*¹, and *Wen Zou*¹

¹Division of Bioinformatics and Biostatistics, National Center for Toxicological Research, U.S. FDA, Jefferson, Arkansas 72079, USA, ²College of Information Engineering, Xiangtan University, Xiangtan, Hunan Province, China.

*Email address for presenting author:* weizhong.zhao@fda.hhs.gov

Next-generation sequencing (NGS) technology has recently been widely applied in clinical and public health laboratory investigations for pathogen detection and surveillance. *Salmonella* is one of the primary causes of foodborne illness in the United States, hundreds of *Salmonella* strains had been collected from food, clinical and environmental sources and their whole genome sequences were obtained by NGS technology. In this study, we developed and implemented a pipeline for sequence acquisition and genetic diversity analysis from NGS data. It consisted of several steps: reference sequence retrieval and template sequence determination; retrieval of NGS sequence reads of *Salmonella* outbreak isolates; multiple sequence alignment and phylogenetic analysis. The pipeline was applied on *Salmonella fliC* gene, which encodes *Salmonella* phase 1 H antigen and is one of the *Salmonella* serotype determinant genes. The *fliC* reference sequences of 24 *Salmonella* strains of 13 serotypes were retrieved from National Center for Biotechnology Information (NCBI) database, and the phylogenetic tree revealed the relationships among the 13 serotypes based on SNPs variations in the data set. The genetic diversity of *Salmonella fliC* gene was distinguished by applying the pipeline on the NGS reads of 48 *S*. Newport, 48 *S*. Montevideo, and 115 *S*. Enteritidis outbreak isolates, respectively. The marker sequences for *Salmonella flic* gene were identified. The developed pipeline provides an effective bioinformatics tool for genetic diversity clarification and marker sequences discovery which will enhance the NGS data analysis and its applications on pathogen identification, source tracking, and population genome evolution.

*Keywords:* Next-generation sequencing, analysis pipeline, gene diversity, *Salmonella* serotypes, *flic* gene