Program

The 4th Workshop on Biostatistics and Bioinformatics

Department of Mathematics and Statistics

Georgia State University

May 8-10, 2015

Contents

Sponsor

The National Science Foundation (NSF), Institute of Mathematical Statistics (IMS), International Chinese Statistical Association (ICSA), Taylor & Francis Group, International Press of Boston, and the Department of Mathematics and Statistics in the Georgia State University

Organizer

Yichuan Zhao

Department of Mathematics and Statistics

Georgia State University

Keynote Speaker

Xiao-Li Meng, Harvard University

Invited Speaker

Michael Epstein, Emory University

Yijuan Hu, Emory University

Liang Liu, University of Georgia

Wenbin Lu, North Carolina State University

Robert Lyles, Emory University

Chris McMahan, Clemson University

Bin Nan, University of Michigan

Annie Qu, University of Illinois

Yiyuan She, Florida State University

Dylan Small, University of Pennsylvania

Wei Wu, Florida State University

Ying Xu, University of Georgia

Daowen Zhang, North Carolina State University

Jiajia Zhang, University of South Carolina

Wenxuan Zhong, University of Georgia

Acknowledgements

The organizer thank Earnestine Collier-Jones, Sandra Ahuama-Jonas, Yvonne Pierce , Leslie Meadows, Tu Tran, Beth Conner, 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 8, 2015

2:00-5:30 pm **Registration**: on the 7th Floor, College of Education Building, 30 Pryor Street.

Saturday, May 9, 2015

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

8:35-8:45 am **Conference Welcome**: Yichuan Zhao, Georgia State University

Opening Remarks: Guantao Chen, Chair of the dept., Georgia State University

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

Personalized Treatment: Sounds heavenly, but where on Earth did they find the right guinea pig for me?

Xiao-Li Meng, Harvard University

9:45-10:00 am **Break:** Refreshments

10:00-11:50 am **Session 2: Chair:** *Jing Zhang , Georgia State University*

Strong Control of the Family-wise Error Rate in Observational Studies

Dylan Small, University of Pennsylvania

Personalized treatment for longitudinal data

Annie Qu, University of Illinois at Urbana-Champaign

Generalized Linear Models with Longitudinal Covariates Subject to Detection Limits

Daowen Zhang, North Carolina State University

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

2:00-3:45 pm **Session 3:** **Chair:** *Wenbin Lu, North Carolina State University*

Large covariance/correlation matrix estimation for temporal data

Bin Nan, University of Michigan

Understanding cancer in its full complexity through

mining cancer tissue omic data

Ying Xu, University of Georgia

Robust Reduced Rank Regression

Yiyuan She, Florida State University

3:45-4:00 pm **Break:** Refreshments

4:00-5:40 pm **Session 4:** **Chair:** *Marko Samara, Georgia State University*

Estimation of Optimal Treatment Regimes for Survival Endpoints from a Classification Perspective

Wenbin Lu, North Carolina State University

Spatial Extended Hazard Model with Application to South Carolina Prostate Cancer Data

*Jiajia Zhang***,** *University of South Carolina*

A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data

Christopher Mcmahan, Clemson 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 10, 2015

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:** *Daowen Zhang, North Carolina State University*

An Efficient Design Strategy for Logistic Regression Using Outcome- and Covariate-Dependent Pooling of Biospecimens Prior to Assay

Robert (Bob) Lyles, Emory University

On Synergy Between Statistical Shape Analysis and Functional Data Analysis

*Wei Wu**,** Florida State University*

A probabilistic model for gene family evolution

*Liang Liu**,** University of Georgia*

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

10:30-12:05pm **Session 6:** **Chair:** *Bin Nan, University of Michigan*

Statistical Approaches for Testing Rare Variants in Affected Pedigrees

Michael Epstein, Emory University

A Reference-free metagenome tool for simultaneously taxonomic classification and distributional estimation of microbial species

Wenxuan Zhong, University of Georgia

Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations

Yijuan Hu, Emory University

12:05-12:10pm **Final Remarks** by Yichuan Zhao

Keynote Talk

Personalized Treatment:
Sounds heavenly, but where on Earth did they find the right guinea pig for me?

Xiao-Li Meng

Department of Statistics, Harvard University

What data are relevant when making a treatment decision for *me*? What replications are relevant for quantifying the uncertainty of this personalized decision? What does "relevant" even mean here? The multi-resolution (MR) perspective from the wavelets literature provides a convenient theoretical framework for contemplating such questions. Within the MR framework, signal and noise are two sides of the same coin: variation. They differ only in the resolution of that variation—a threshold, the *primary* resolution, divides them. We use observed variations at or below the primary resolution (signal) to estimate a model and those above the primary resolution (noise) to estimate our uncertainty. The higher the primary resolution, the more relevant our model is for predicting a personalized response. The search for the appropriate primary resolution is a quest for an age old bias-variance trade-off: estimating more precisely a less relevant treatment decision versus estimating less precisely a more relevant one. However, the MR setup crystallizes how the tradeoff depends on three objects: (i) the estimand which is independent of any statistical model, (ii) a model which links the estimand to the data, and (iii) the estimator of the model. This trivial, yet often overlooked distinction, between estimand, model, and estimator, supplies surprising new ways to improve mean squared error. The MR framework also permits a conceptual journey into the counterfactual world as the resolution level approaches infinite, where "me" becomes unique and hence can only be given a single treatment, necessitating the potential outcome setup. A real-life Simpson's paradox involving two kidney stone treatments will be used to illustrate these points and engage the audience.

This talk is based on the following three articles:

Meng, X.-L. (2014): A Trio of Inference Problems that Could Win You a Nobel Prize in Statistics (If You Help Fund It). *In Past, Present, and Future of Statistical Science (Eds Lin et. al.).* (Available at http://www.stat.harvard.edu/Faculty_Content/meng/COPSS_50.pdf)

Liu, K and Meng, X.-L. (2014): A Fruitful Resolution to Simpson's Paradox via Multi-Resolution Inference. *The American Statistician, Vol* 68, pp 17-29. (available at http://statistics.fas.harvard.edu/people/xiao-li-meng)

Liu, K. and Meng, X.-L. (2015). There is individualized treatment. Why not individualized inference? To appear in *Annual Review of Statistics and Its Applications*. (available by emailing meng@stat.harvard.edu).

Invited Talks

Statistical Approaches for Testing Rare Variants in Affected Pedigrees

Michael Epstein

Emory University

While many rare-variant association tests of disease exist for case-control designs, far fewer analogous methods exist for affected pedigrees. This is unfortunate, since affected pedigrees have many attractive features for rare-variant analysis that case-control studies lack. Many studies are sequencing familial samples, particularly those collected from past linkage projects. To allow efficient analysis in such pedigrees, we propose a statistical framework for rare-variant association testing in families based on the idea that rare susceptibility variants should be found more on regions shared identical by descent by affected relative pairs than on regions not shared identical by descent. Using simulated data, we show our approach for rare-variant association testing in affected pedigrees is more powerful than standard case-control association testing assuming fixed sample size and our approach also has the additional benefit of being robust to confounding due to population stratification. We illustrate the approach using rare and less-common variation from a family-based study of hypertension.

This represents joint work with Dr. Glen Satten at the CDC.

Integrative analysis of sequencing and array genotype data for discovering disease associations with rare mutations

Yijuan Hu

Emory University

In the large cohorts that have been used for genome-wide association studies (GWAS), it is prohibitively expensive to sequence all cohort members. A cost-effective strategy is to sequence subjects with extreme values of quantitative traits or those with specific diseases. By imputing the sequencing data from the GWAS data for the cohort members who are not selected for sequencing, one can dramatically increase the number of subjects with information on rare variants. However, ignoring the uncertainties of imputed rare variants in downstream association analysis will inflate the type I error when sequenced subjects are not a random subset of the GWAS subjects. In this article, we provide a valid and efficient approach to combining observed and imputed data on rare variants. We consider all commonly used genelevel association tests, all of which are based on the score statistic for assessing the effects of individual variants on the trait of interest. We show that the score statistic based on the observed genotypes for sequenced subjects and the imputed genotypes for non-sequenced subjects is unbiased. We derive a robust variance estimator that reflects the true variability of the score statistic regardless of the sampling scheme and imputation quality, such that the corresponding association tests always have correct type I error. We demonstrate through extensive simulation studies that the proposed tests are substantially more powerful than the use of accurately imputed variants only and the use of sequencing data alone. We provide an application to the Women's Health Initiative (WHI). The relevant software is freely available.

A probabilistic model for gene family evolution

Liang Liu

University of Georgia

Gene duplication played a pivotal role in evolution (Ohta 1989). The widespread existence of gene families suggests that the newly arisen gene duplicates are the major contributors to evolutionary novelties (Lynch et al. 2001; Hurles 2004). The extra gene copies resulted from duplication provided raw genetic material that evolutionary forces can act on. Although a majority of duplicate genes may be silenced by degenerative mutations, some duplicated genes are able to evolve novel functions permanently preserved in the population (Lynch 2002). Accurately estimating the timing and mode of gene duplications along the evolutionary history of species can provide invaluable information about underlying mechanisms by which the genomes of organisms evolved and the genes with novel functions arose (Hahn et al. 2005). As genomic data become increasingly available, it is desirable to build a model based on a stochastic process that is a good approximation to the real biological process of gene duplication and loss.

Such probabilistic model can both add biological realism to improve the fit of the model to the data as well as enable mechanistic inference that is currently not possible.

An Efficient Design Strategy for Logistic Regression Using Outcome- and Covariate-Dependent Pooling of Biospecimens Prior to Assay

Robert (Bob) Lyles

Emory University

Potential reductions in laboratory assay costs afforded by pooling equal aliquots of biospecimens have long been recognized in disease surveillance and epidemiological research and, more recently, have motivated design and analytic developments in regression settings. For example, Weinberg and Umbach (1999, *Biometrics* **55,** 718-726) provided methods for fitting set-based logistic regression models to case-control data when a continuous exposure variable (e.g., a biomarker) is assayed on pooled specimens. We focus on improving estimation efficiency by utilizing available subject-specific information at the pool allocation stage. We find that a strategy that we call "(y,**c**)-pooling," which forms pooling sets of individuals within strata defined jointly by the outcome and other covariates, provides more precise estimation of the risk parameters associated with those covariates than does pooling within strata defined only by the outcome. We review the approach to set-based analysis through offsets developed by Weinberg and Umbach, along with a recent correction to their original paper. We also consider an alternative approach to set-based analysis based on offsets derived as simple functions of inverse probability weights. We propose a method for variance estimation under this design and use simulations and a real-data example to illustrate the precision benefits of (y,**c**)-pooling relative to y-pooling. We also note and illustrate that set-based models permit estimation of covariate interactions with exposure.

Estimation of Optimal Treatment Regimes for Survival Endpoints from a Classification Perspective

Wenbin Lu

North Carolina State University

A treatment regime is a deterministic function that dictates personalized treatment based on patients' individual prognostic information. There is a fast-growing interest in finding optimal treatment regimes to maximize expected long-term clinical outcomes of patients for complex diseases. For many clinical studies with survival time as a primary endpoint, a main goal is to maximize patients' survival probabilities given treatments. In this talk, I will present a doubly robust estimator for the value function with censored survival data and study the large sample properties of this estimator. The optimization for searching the optimal treatment regime is realized from a weighted classification perspective that allows us to use available off the shelf software. In some studies one treatment may have greater toxicity or side effects, thus we also consider estimating a quality adjusted optimal treatment regime that allows a patient to trade some additional risk of death in order to avoid the more invasive treatment. The proposed methods were used to estimate optimal treatment regimes in the ASCERT study of patients with coronary artery disease.

A flexible, computationally efficient method for fitting the proportional hazards model to interval-censored data

Christopher Mcmahan

Clemson University

The proportional hazards model (PH) is currently the most popular regression model for analyzing time-to-event data. Despite its popularity, the analysis of interval-censored data under the PH model can be challenging using many available techniques. A new method for analyzing interval-censored data under the PH model will be presented. The proposed approach uses a monotone spline representation to approximate the unknown nondecreasing cumulative baseline hazard function. Formulating the PH model in this fashion results in a finite number of parameters to estimate while maintaining substantial modeling flexibility. A novel expectation-maximization (EM) algorithm is developed for finding the maximum likelihood estimates of the parameters. The derivation of the EM algorithm relies on a two-stage data augmentation involving latent Poisson random variables. The resulting algorithm is easy to implement, robust to initialization, enjoys quick convergence, and provides closed-form variance estimates. The performance of the proposed regression methodology is evaluated through a simulation study, and is further illustrated using data from a large population-based randomized trial designed and sponsored by the United States National Cancer Institute.

Large covariance/correlation matrix estimation for temporal data

Bin Nan

University of Michigan

We consider the estimation of high-dimensional covariance and correlation matrices under slow-decaying temporal dependence. For generalized thresholding estimators, convergence rates are obtained and properties of sparsistency and sign-consistency are established. The impact of temporal dependence on convergence rates is also investigated. An intuitive cross-validation method is proposed for the thresholding parameter selection, which shows good performance in simulations. Convergence rates are also obtained for banding method if the covaraince or correlation matrix is bandable. The considered temporal dependence has longer memory than those in the current literature and has particular implications in analyzing resting-state fMRI data for brain connectivity studies.

Personalized treatment for longitudinal data

Annie Qu

University of Illinois at Urbana-Champaign

We develop new modeling and estimation for personalized treatment for individuals with high heterogeneity. Incorporating subject-specific information into treatment subgroup is critical since individuals could react to the same treatment quite differently. We propose to identify subgroups with longitudinal observations through random-effects estimation where the random effects are not necessarily normal distributed. The advantage of this approach is that we can quantify intrinsic associations between unobserved subject-specific effects and observed treatment outcomes, and therefore provide optimal treatment assignments for different individuals. In contrast, traditional mixed-effects models assuming normal distribution cannot effectively distinguish different patterns of treatment effects. We develop asymptotic consistency theory for individual treatment effect estimation, and show that the new estimator is more efficient than the random effect estimator which ignores correlation information from longitudinal data. Simulation studies and a data example from an AIDS clinical trial group confirm that the proposed method is quite efficient in identifying an effective treatment strategy for subgroups in finite samples.

This is joint work with Hyunkeun Cho and Peng Wang.

Robust Reduced Rank Regression

Yiyuan She

Florida State University

This work is motivated by Arabidopsis Thaliana data analysis where we want to find the relationship between the genes from MVA and MEP and the genes from some downstream pathways (plastoquinone, caroteniod, phytosterol and chlorophyll). In such multivariate problems, rank reduction is a very effective way for dimension reduction to facilitate model estimation and interpretation. However, commonly used reduced rank methods are extremely non-robust against data corruption, as the low-rank dependence structure between response variables and predictors could be easily distorted in the presence of gross outliers. We propose a robust reduced rank regression approach for joint reduced rank modeling and outlier detection. The proposed approach generalizes and unifies several popular robust estimation methods. Theoretically, nonasymptotic error bounds are developed, which demonstrate that conducting rank reduction and outlier detection jointly leads to improved prediction accuracy. The performance of the proposed method is examined empirically by simulation studies and real data examples.

This is joint work with Kun Chen.

Strong Control of the Family-wise Error Rate in Observational Studies

Dylan Small
University of Pennsylvania, Department of Statistics

An effect modifier is a pretreatment covariate such that the magnitude of the treatment effect or its stability changes with the level of the covariate. Generally, other things being equal, larger treatment effects and less heterogeneous treatment effects are less sensitive to unmeasured biases in observational studies. It is known that when there is effect modification, an overall test that ignores an effect modifier may report greater sensitivity to unmeasured bias than a test that combines results at different levels of the effect modifier. This known combined test reports that there is evidence of an effect somewhere that is insensitive to bias of a certain magnitude, but it does not draw inferences about affected subgroups. If there is effect modification, one would like to identify specific subgroups for which there is evidence of effect that is insensitive to small or moderate biases. We propose an exploratory method for discovering effect modification combined with a confirmatory method of simultaneous inference that strongly controls the family-wise error rate in a sensitivity analysis, despite the fact that the groups being compared are defined empirically. A new form of matching, strength k matching, permits a search through more than k covariates for effect modifiers, yet no pairs are lost providing at most k covariates are selected to group the pairs. In a strength k match, each set of k covariates is exactly balanced, though a set of > k covariates may exhibit imbalance. We apply the method to study the effects of the powerful earthquake that struck Chile in 2010.

This is joint work with Jesse Hsu, Jose Zubizarreta and Paul Rosenbaum.

On Synergy Between Statistical Shape Analysis and Functional Data Analysis

Wei Wu

Florida State University

The problem of statistical shape analysis (SSA) of objects has traditionally been formulated as an analysis of landmarks (registered points) modulo certain similarity transformations (rotation, translation, and scale). More recently SSA techniques have been extended to include shapes of continuous objects -- parameterized curves, surfaces, and their temporal evolutions -- by treating them as elements of Hilbert spaces. The branch of statistics on functional data functional data analysis (FDA) -- also deals with generating inferences on certain Hilbert spaces and shares some common issues and solutions with SSA. Specifically, the problem of phase-amplitude separation in FDA involves alignment of peaks and valleys of given functions using nonlinear time warpings. This relates to optimal registration of points in domains of the given functions using diffeomorphism maps. This exact problem has been termed as the registration in SSA. An elegant solution to this problem comes from use of a family of square-root transforms of the original functions or objects, along with the standard L2 norm. I will describe this framework using examples from FDA and SSA of curves and surfaces.

Understanding cancer in its full complexity through mining cancer tissue omic data

Ying Xu

Department of Biochemistry & Molecular Biology

Computational Systems Biology Lab

University of Georgia

A vast majority of the published cancer studies in the past few decades was conducted on cancer cells rather than cancer tissues. Knowing that the microenvironment plays key roles in cancer initiation, development and metastasis, we must reassess the true relevance of many of these published results to cancer. We have recently developed a new framework for cancer studies by treating cancer as a survival process under increasingly more challenging stresses, which evolve as a cancer evolves. Our main hypothesis is that cell proliferation is a sustained and common pathway to survival under all major cancer-associated stresses. The availability of large-scale cancer tissue omic data enables us to systematically identify various stress types present in each tissue and how each cancer tumor responds to the encountered stresses, ultimately validating, refining or rejecting this fundamentally novel hypothesis. In this presentation, I will discuss (1) how data mining can be used to identify such stresses and their responses, leading to substantially improved understanding about cancer evolution from its onset; and (2) how data mining-based discoveries can be integrated with cell-based experimental findings, leading to more comprehensive understanding about the key drivers and facilitators of cancer evolution, hence potentially leading to much improved treatment paradigms for challenging cancer cases.

Generalized Linear Models with Longitudinal Covariates Subject to Detection Limits

Daowen Zhang
Department of Statistics
North Carolina State University

In the motivating GenIMS study, it it of public health interest to investigate the effects of longitudinal covariates subject to detection limits on the 90-day survival of patients with community acquired pneumonia (ACP). We considered GLMs with random effects as covariates from the mixed models for longitudinal covariates to address the research objective. Likelihood inference is complicated due to the large number of random effects and the detection limit issue. We proposed a fast, approximate EM algorithm that reduces the dimension of integration in the E-step of the algorithm to one, regardless of the number of random effects in the joint model. Numerical studies demonstrate that the proposed approximate EM algorithm leads to satisfactory parameter and variance estimates in situations with and without censoring on the longitudinal covariates. The approximate EM algorithm is applied to analyze the GenIMS data set, yielding meaningful results.

Joint work with Paul Bernhardt Villanova University and Huixia Judy Wang George Washington University

Keywords: Detection limit, EM algorithm, Joint model, Logistic regression

Spatial Extended Hazard Model with Application to South Carolina Prostate Cancer Data

Jiajia Zhang

University of South Carolina

This project quantifies racial cancer survival disparity in South Carolina through the consideration of a Bayesian semiparametric approach to the extended hazards model, with generalization to high-dimensional spatially-grouped data. The baseline hazard function is modeled using a novel penalized B-spline that a priori follows a parametric hazard function. County-level spatial correlation is accommodated marginally through the copula model of Li and Lin (2006), using a correlation structure implied by an intrinsic conditionally autoregressive prior. Efficient MCMC algorithms are developed, especially applicable to fitting very large, highly-censored areal survival data sets. Per-variable tests for proportional hazards, accelerated failure time, and accelerated hazards are efficiently carried out with and without spatial correlation through Bayes factors. The resulting reduced, highly interpretable spatial models fit significantly better than the additive Cox model with spatial frailties.

A Reference-free metagenome tool for simultaneously taxonomic classification and distributional estimation of microbial species

Wenxuan *Zhong*

University of Georgia

MetaGenomics refers to the study of a collection of genomes, typically microbial genomes, presenting in environmental samples, such as samples from the gastrointestinal tract of a human patient or samples of soil from a particular ecological origin. By sequencing bulk DNA that is directly extracted from environmental samples, one can bypass the difficulties arising in cell cultivation. Moreover, we can easily identify novel microbial species and study their distribution variation along different samples. However, these advantages cannot really benefit the biological researcher before we have a high-resolution and reference-free MetaGenomic tool, because most existing methods that focus on basic taxonomy ranks need to align short reads to a reference genome. It is very challenging to identify species that are not well studied. The method we introduce in this article can overcome this difficulty and provide a reliable detection of microbial species. Our method leverages the matrix factorization method to simultaneously estimating known and unknown species and their proportions in a microbial colony. We demonstrate our method in both simulation and a real biological study.

Poster Abstracts

Empirical Likelihood Confidence Intervals for the Population Mean Based on Incomplete Data

Jose Valdovinos Alvarez

Georgia State University

The use of doubly robust estimators is a key for estimating the population mean response in the presence of incomplete data. Cao et al. (2009) proposed an alternative doubly robust estimator which exhibits strong performance compared to existing estimation methods. In this paper, we apply the jackknife empirical likelihood, the jackknife empirical likelihood with nuisance parameters, the pro file empirical likelihood, and an empirical likelihood method based on the influence function to make an inference for the population mean. We use these methods to construct confidence intervals for the population mean, and compare the coverage probabilities and interval lengths using both the "usual" doubly robust estimator and the alternative estimator proposed by Cao et al. (2009). An extensive simulation study is carried out to compare the different methods. Finally, the proposed methods are applied to two real data sets.

Joint work with Yichuan Zhao.

Jackknife empirical likelihood inference for the difference of two volumes under the ROC surfaces

Yueheng An and Yichuan Zhao

Georgia State University

The volume under a surface, abbreviated VUS, is an effective method for evaluating the discriminating power of a diagnostic test with three ordinal diagnostic groups. In this paper, the difference of two VUS's is investigated to compare two treatments for the discrimination of three-category classification data. A jackknife empirical likelihood (JEL) procedure is employed to avoid the estimation of several simultaneous equations in the existing methods. We prove that

the limiting distribution of the log empirical likelihood ratio statistic follows chi-squared distribution. Numerical studies show that the JEL based confidence intervals outperform those based on the normal approximation method in terms of shorter average length and more precise coverage probability for small and moderate sample sizes.

Variable selection under signed-rank regression

Frazier Bindele

University of South Alabama

The growing need for dealing with big data has made it necessary to find computationally efficient methods for identifying important factors to be considered in statistical modeling. In the linear model, the LASSO is an effective way of selecting variables using penalized regression. It has spawned substantial research in the area of variable selection for models that depend on a linear combination of predictors. However, with the exception of a few instances, there are has not been much work addressing the lack of optimality of variable selection when the model errors are not Gaussian and/or when the data contain gross outliers. We propose the signed-rank LASSO as a robust and efficient alternative to LAD and LS LASSO. As LAD LASSO, the approach is appealing for use with big data since one can use data augmentation to perform the estimation as a single weighted L1 optimization problem.

This is based on joint work with Ash Abebe (Auburn University).

Extending Semiparametric AUC Model with Discrete Covariates into General Framework

*Som B. Bohora, MS; *Yan D. Zhao, PhD; and Tatiana N. Balachova, PhD*

The University of Oklahoma Health Sciences Center, OK 73104, USA

In the context of analyzing non-normally distributed data, an existing AUC model that handles only two discrete covariates was extended to a general framework that can adjust for any number of categorical covariates with any levels and their interactions with the treatment group. Compared to other similar methods which require iterative algorithms and bootstrap procedure, our method involved only closed-form formulae for parameter estimation, hypothesis testing, and confidence intervals. The issue of model identifiability was also discussed. Our model has broad applicability in clinical trials due to the ease of interpretation on model parameters and its utility was illustrated using data from a clinical trial study aimed at evaluating education materials for prevention of Fetal Alcohol Spectrum Disorders (FASDs). Finally, for a variety of design scenarios, our method produced parameters with small biases and confidence intervals with nominal coverages as demonstrated by simulations.

Keywords: non-normal response, semiparametric AUC, discrete covariates, generalization, clinical trial

Regulation of Ionic Dynamics by the β_1-Adrenergic Signaling in Mouse Ventricular Myocytes

Vladimir E. Bondarenko

Department of Mathematics and Statistics and Neuroscience Institute,

Georgia State University, Atlanta GA, USA

The β_1-adrenergic signaling system is one of the most important protein signaling systems in cardiac cells. It regulates cardiac electrical activity and intracellular ionic concentration. Dysfunction of the β_1-adrenergic signaling system results in cardiac hypertrophy, which leads to heart failure. In this presentation, a comprehensive, experimentally-based mathematical model of the β_1-adrenergic signaling system for mouse ventricular myocytes is explored to simulate the effects of relatively moderate stimulations of β_1-adrenergic receptors on Ca^{2+} and Na^+ dynamics, as well as the effects of inhibition of protein kinase A and phosphodiesterase of type 4. Simulation results show the differences in the response of Ca^{2+} and Na^+ fluxes to moderate stimulation of β_1-adrenergic receptors. The investigated model reproduced most of the experimentally observed effects of PKA and PDE4 inhibition on the L-type Ca^{2+} current, $[Ca^{2+}]_i$ transients, and the sarcoplasmic reticulum Ca^{2+} load, and made testable predictions for the action potential duration and $[Ca^{2+}]_i$ transients as functions of isoproterenol concentration.

Quantifying Uncertainty in the Identification of Proteins, Post-translational Modifications (PTMs) and Proteoforms

Naomi Brownstein

Florida State University and National High Magnetic Field Lab

The traditional goals of top-down proteomics are protein identification and quantitation. However, the presence of additional sources of variability, such as post-translational modifications (PTMs) and genomic variants, complicates the problem of identification. Recent interest in the proteomics community has begun to shift from the relatively narrow problem of protein identification to consideration of these additional sources of variability. Combining these factors results in a unique exhaustively defined chemical species termed 'proteoform.' We explore a variety of scoring metrics and estimate their uncertainty via bootstrapping. We demonstrate the method using human histone H4 and the corresponding proteoforms. Results show that related proteoforms may be statistically difficult to differentiate.

This work was supported by NSF Division of Materials Research through DMR-11-5749 and the State of Florida.

LOGIC REGRESSION IN THE BIO-MEDICAL SCIENCES:

METHODOLOGY AND APPLICATIONS

Sujay Datta
Dept. of Statistics
University of Akron

In the bio-medical sciences, complex interactions among variables or factors often play an important role in knowledge discovery, provide crucial insights into the underlying biology and influence practice patterns. Unfortunately, under conventional regression models, anything more complicated than three-way interactions is rarely included. When most of the variables or factors are binary, LOGIC REGRESSION provides a convenient way of incorporating higher order interactions using Boolean logic operators. In this brief overview, we describe a logic regression model, provide some theoretical details of how it works and finally give three examples of its use in the contexts of identifying interacting SNPs (single nucleotide polymorphisms), discovering regulatory motifs and assessing provider effects on kidney cancer treatment delivery.

Joint work with Bing Liu.

Analyze mRNA Sequencing Data Using Empirical Bayes Selection Method

Cuilan (Lani) Gao

University of Tennessee at Chattanooga

Differential expression analysis of sequencing count expression data involves performing a large number of hypothesis tests that compare the expression count data of each gene or transcript across two or more biological conditions. The assumptions of any specific hypothesis-testing method will probably not be valid for each of a very large number of genes. Thus, computational evaluation of assumptions should be incorporated into the analysis to select an appropriate hypothesis testing method for each gene. Here, we generalize earlier work to introduce two novel procedures that use estimates of the empirical Bayesian probability (EBP) of over dispersion to select or combine results of a standard Poisson likelihood ratio test and a quasi-likelihood test for each gene. These EBP-based procedures simultaneously evaluate the Poisson-distribution assumption and account for multiple testing. The new procedures select the standard likelihood test for each gene with Poisson-distributed counts and the quasi-likelihood test for each gene with overdispersed counts. The new procedures outperformed previously published methods in many simulation studies.

Classification of mouse RPE by K-nearest neighbor and logistic regression

Haitao Huang

Georgia State University

In age-related macula degeneration (AMD), retinal pigment epithelium (RPE) cells undergo cell apoptosis and cell deformation. Morphometric measures of RPE cells can be used to discriminate the age and disease status of subjects. Using the K-nearest neighbor algorithm and logistic regression, we classified the genotypes and ages by RPE cell morphometric measures. Our result suggest there is little difference in prediction rates in angular locations, but significant differences in radial locations. Variables such as eccentricity and extent consistently have prediction accuracy as high as 90%.

An ensemble method of k-mer and Natural Vector for multiple-segmented virus classification

Hsin-Hsiung Huang

University of Central Florida

The Natural Vector combined with Hausdorff distance has been successfully applied for classifying and clustering multiple-segmented viruses. Additionally, *k*-mer methods also yield promising results for global genome comparison. It is not known whether combining these two approaches can lead to more accurate results. The author proposes a combination of the Hausdorff distances of the 5-mer counting vectors and natural vectors which achieves the best classification without cutting off any sample. Using the proposed method to predict the taxonomic labels for the 2,363 NCBI reference viral genomes dataset, the accuracy rates are 96.95%, 94.37%, 99.41% and 93.82% for the Baltimore, family, subfamily, and genus labels, respectively. We further applied the proposed method to 68 isolates of the influenza A H7N9 viruses which consist of eight segments. The resulting natural graphical representations show that the proposed method can lead to the most reasonable phylogenetic relationships.

Semiparametric mixed model analysis for nonlinear gene-environment interactions in genome-wide association study

Zijian Huang

University of California, Riverside

The use of linear mixed models (LMMs) in genome-wide association studies (GWAS) is now widely accepted because LMMs have been shown to be capable of correcting for genetic relatedness of sampled data. On the other hand, gene and environment (G × E) interactions play a pivotal role in determining the risk of human diseases. Conventional parametric models such as linear mixed model may not reflect the underlying nonlinear G × E interaction, which will result in serious bias. In this paper, we propose a semiparametric mixed model to investigate important

gene associations in the context of possible nonlinear G × E interactions in GWAS. We further propose a profile maximum likelihood estimation procedure to estimate the parametric and nonparametric functions, and apply quasi restricted maximum likelihood estimation method to estimate the variance components. For these profile parameter and nonparametric function estimators, asymptotic consistency and normality are established. Moreover, Rao-score-type test procedure is developed to identify the important genetic factors. Both simulation studies and an empirical example are presented to illustrate the use of our proposed model and methods.

Coauthorship and Citation Networks for Statisticians

Pengsheng Ji

University of Georgia

We collect the coauthor and citation data for all research papers published in four of the top journals in statistics between 2003 and 2012, analyze the data from several different perspectives. We identify some hot areas and most cited authors and papers, and also a handful of meaningful communities, such as "high-dimensional data", "large-scale multiple testing", "Dimensional Reduction", "Objective Bayes" and "Theoretical Machine Learning", etc.

This is a joint work with Jiashun Jin at Carnegie Mellon.

Sample Size Determination of a Hierarchal Designed Model in MCM2 Biomarker in Normal and Premalignant Tissue Associated with Prostate Cancer

Nivedita Kar

Northwestern University

Statistical consideration in clinical research does not sufficiently address the issue of sub-sampling in hierarchal models where multiple levels of measurements are being sampled (1). Through a model discussed by Jovanovic et al. (1), the number of needed units in all tiers of a hierarchal investigation can be calculated as a function of variance to optimize subsequent sampling error and cost. An application motivated by prostate cancer presents a hierarchal designed experiment in which each of 10 participants prostate gland measures 6 different intraepithelial tissue compartments and each compartment is measured in 10 sites on the prostate gland for luminal expression of a biomarker, Minichromosome Maintenance Protein-2 (MCM-2) (2). The objective of this paper is to adapt the model to calculate an optimal sample size with consideration for sub-sampling measurements for a potential RCT to reduce the expression of the MCM-2 biomarker in the normal tissue compartment. Three calculated subject level sample sizes will be presented with regard for differing scenarios of variant numbers of site measurements being sampled. These scenarios will be represented through 3 calibrated inflation ratios or the percentage increase needed on subject-level due to variance and sampling at sub-subject level. This methodology is advisable to be employed in the design phase of nested clinical research experimentation.

References:

1. Jovanovic B, Subramanian H, Helenowski I, et al. Northwestern University, Clinical trial laboratory data nested within subject: components of variance, sample size, and cost. “Unpublished Manuscript”
2. Ananthanarayanan V, Deaton R, Yang X, et al. Alteration of proliferation and apoptotic markers in normal and premalignant tissue associated with prostate cancer. BMC cancer. 2006; 6(1): 73.

A New Approach for Detecting CNV regions using the Next Generation Sequencing Reads

Jaeeun Lee

Department of Biostatistics and Epidemiology

Medical College of Georgia, Georgia Regents University

Modeling the high-throughput next generation sequencing (NGS) data, resulting from experiments with the goal of profiling tumor and control samples for the study of DNA copy number variants (CNVs), remains to be a challenge in various ways. In this work, we provide an efficient method for detecting multiple CNVs using NGS reads ratio data. This method is based on a multiple statistical change-points model with the 1d fused lasso (least absolute shrinkage and selection operator) that is designed for ordered data in a one-dimensional structure. Since the absolute penalty in the 1d fused lasso gives a piecewise constant solution and enforces sparsity of the coefficients' successive differences, the 1d fused lasso is suitable for detecting change points. In addition, since the path algorithm traces the solution as a function of a tuning parameter, the number and locations of potential CNV region boundaries can be identified in an efficient way where the ratios of NGS reads are also estimated simultaneously. For tuning parameter selection, we then propose a new modified Bayesian information criterion, called the JMIC, and compare the JMIC with three different Bayes information criteria: SIC, PMIC, and ZMIC. We applied our approach to the sequencing data of reads ratio between the breast tumor cell lines HCC1954 and its matched normal cell line BL 1954 on chromosomes 1 and 8 and the results show better performance of JMIC for tuning parameter selection, in comparison with SIC, PMIC, and ZMIC.

This is a joint work with Dr. Jie Chen.

Mixtures of g-priors in Generalized Linear Models

Yingbo Li

Clemson University

Mixtures of Zellner's g-priors have been studied extensively in linear models and have been shown to have numerous desirable properties for Bayesian variable selection and model averaging. Several extensions of g-priors to Generalized Linear Models (GLMs) have been proposed in the literature; however, the choice of prior distribution of g and resulting properties for inference have received considerably less attention. In this paper, we extend mixtures of g-priors to GLMs by assigning the truncated Compound Confluent Hypergeometric (tCCH) distribution to 1/(1 + g) and illustrate how this prior distribution encompasses several special cases of mixtures of g-priors in the literature, such as the Hyper-g, truncated Gamma, Beta-prime, and the Robust prior. Under an integrated Laplace approximation to the likelihood, the posterior distribution of 1/(1 + g) is in turn a tCCH distribution, and approximate marginal likelihoods are thus available analytically. We discuss the local geometric properties of the g-prior in GLMs and show that specific choices of the hyper-parameters satisfy the various desiderata for model selection proposed by Bayarri et al [2012], such as asymptotic model

selection consistency, information consistency, intrinsic consistency, and measurement invariance. We also illustrate inference using these priors and contrast them to others in the literature via simulation and real examples. (co-authored with Merlise Clyde)

Identifying Time Windows of Susceptibility to Time-varying Exposures of Complex Metal Mixtures

Shelley Han Liu

Harvard University

Exposures to heavy metal mixtures (ie. lead, manganese, zinc, cadmium, etc.) may significantly impact neurodevelopment in early life. Due to sequential neurodevelopmental processes, there may be certain time windows of susceptibility during which vulnerability to metal mixtures is increased; studies have shown effect modification of one metal's exposure in the presence of other metals, and a possibility for detectable mixture effects on health at low doses of exposure below individual no observable adverse effect levels. We present a Bayesian hierarchical modeling framework that identifies time windows of susceptibility in the context of exposure to metal mixtures. This method combines Bayesian penalization through the group and fused lasso with Bayesian kernel machine regression to flexibly capture exposure-response relationships of metal mixtures, incorporate prior knowledge, account for collinearity of mixture components, and account for non-linear and non-additive effects of individual exposures to identify sensitive time windows. Simulations demonstrate the ability of the method to detect time-varying nonlinear and quadratic effects, and the method is applied to a study of heavy metal mixture exposure in children.

A flexible procedure to build generalized additive partially linear models for high-dimensional data, with an application to analyze gene expression data in a breast cancer study

Xiang Liu

Health Informatics Institute

University of South Florida

In cancer and other medical research, the mechanism of the disease is complicated. Nowadays, datasets involving a large number of measurements (such as gene expression data and other -omic data) are produced with the hope to discover the relationship between the measurements and the phenotype. Establishing a viable model to explore the relationship between high-dimensional covariates and binary responses can be quite challenging. One challenge comes from the large number of covariates especially when the number of covariates (*p*) is larger than the number of observations (*n*). Another challenge is that the assumption of a linear relationship between the covariates and log-odds-ratio in logistic regression (which is commonly used for binary response) may not fit the data well. In order to address these two challenges, we propose a flexible procedure of building models to analyze high-dimensional data. Generalized additive partially linear models (GAPLM) are used to model the effects of covariates on the responses, which can reflect nonlinear effects of some covariates on the log-odds-ratio. The covariates in the model are determined by a novel variable selection method using bootstrapping and penalized regression. We apply the procedure to analyze gene expression data from a breast cancer study. The example shows that the proposed procedure is very flexible and practically useful. A simulation study is also conducted to demonstrate the good performance of the procedure.

Keywords: Generalized additive partially linear models; LASSO; Bootstrap; High dimensionality; Group selection; Gene expression data

Joint work with Tian Chen, Yuanzhang Li and Hua Liang

A sensitivity analysis of different regression models for competing risks data through a simulation study

Yuliang Liu

University of Alabama at Birmingham

In order to compare the performance of three approaches: Cox PH model, Exponential Accelerated Failure Time (AFT) model and Fine Gray’s model for the estimation of hazard ratio of event of interest for the competing data, we conducted a sensitivity analysis of these models through a simulation study. The competing risks datasets were generated by simulations following latent failure model or bivariate random variables model. For the result of Power, type I error, standard error, we did not see any significant difference among the datasets generated by different simulation methods or between three regression models. However, for the analysis of bias and interval coverage, the Fine-Gray PSH model showed a large bias and lower interval coverage compared to the other two models. When we use the estimate of time-averaged subdistribution hazard ratio of event of interest instead of the true value of the hazard ratio of event of interest to measure the average bias and interval coverage in Fine-Gray PSH model, the difference of results between these models disappear. We confirmed that a subdistribuiton hazards regression model has a proper interpretation, even when the subdistribution hazards were falsely assumed to be proportional.

Jackknife Empirical Likelihood for the Concordance Correlation Coefficient

Anna Susan Moss

Georgia State University

Lin's Concordance Correlation Coefficient (CCC) is a commonly used measure of reproducibility or agreement in paired samples. Interval estimates for the CCC are typically constructed using the normal approximation (NA) method (Lin, 1989) which can have poor coverage for samples from skewed distributions. To improve coverage, we developed nonparametric confidence intervals for the CCC based on the jackknife empirical likelihood (JEL) method (Jing, 2009). We evaluated four JEL methods: the unadjusted (or original) JEL, the adjusted JEL (AJEL), extended JEL (EJEL), and bootstrap-calibrated JEL (BCJEL) for interval estimation of the CCC. We compared performance of the four JEL methods to the NA and two bootstrap methods. We computed coverage probability and average interval length from paired samples generated from normal, exponential, Poisson, uniform distributions and lognormal-normal mixed models. The JEL attained the same coverage as the NA method for normal data and had better coverage than NA and bootstrap methods for non-normal data. Generally the extended and bootstrap-calibrated JEL achieved the best coverage probabilities particularly with small sample of data from highly skewed distributions. The JEL methods produced wider confidence intervals compared to the NA method. We illustrated the application of JEL methods to the CCC using self-reported and clinically measured height and weight from the National Health and Nutrition Examination Survey (NHANES). It is joint work with Yichuan Zhao.

Symmetric Gini Covariance and Correlation

Yongli Sang

Department of Mathematics, University of Mississippi

Gini covariance and Gini correlation play important roles in measuring the dependence of random variables with heavy tails. However, its asymmetry brings a substantial difficulty in interpretation. In this paper, we propose a new symmetric Gini-type covariance based on the joint rank function. As a result, a new correlation called the symmetric Gini correlation is also defined. We study the properties of the symmetric Gini correlation. Its influence function shows that it is more robust than Pearson correlation but less robust than Kendall's correlation. The relationship between the symmetric Gini correlation and the linear correlation is established for a class of random vectors in the family of elliptical distributions. With this relationship, an estimator of the linear correlation is obtained from estimating the Gini correlation. We study the asymptotic normality of this estimator through two approaches: one from influence function and the other from U-statistics and delta method. We compare asymptotic efficiencies of linear correlation estimators based on Gini, Pearson and Kendall's correlation under various distributions. The proposed one not only balances between robustness and efficiency, its superior finite sample behavior also makes it advantageous in application.

Binary regression with differentially misclassified response and exposure variables.

Li Tang

St. Jude Children's Research Hospital

Misclassification is a long-standing statistical problem in epidemiology. In many real studies, either an exposure or a response variable or both may be misclassified. As such, potential threats to the validity of the analytic results (e.g., estimates of odds ratios) that stem from misclassification are widely discussed in the literature. Much of the discussion has been restricted to the nondifferential case, in which misclassification rates for a particular variable are assumed not to depend on other variables. However, complex differential misclassification patterns are common in practice, as we illustrate here using bacterial vaginosis and Trichomoniasis data from the HIV Epidemiology Research Study (HERS). Therefore, clear illustrations of valid and accessible methods that deal with complex misclassification are still in high demand. We formulate a maximum likelihood (ML) framework that allows flexible modeling of misclassification in both the response and a key binary exposure variable, while

adjusting for other covariates via logistic regression. The approach emphasizes the use of internal validation data in order to evaluate the underlying misclassification mechanisms. Data-driven simulations show that the proposed ML analysis outperforms less flexible approaches that fail to appropriately account for complex misclassification patterns. The value and validity of the method are further demonstrated through a comprehensive analysis of the HERS example data.

Joint work with Lyles RH, King CC, Celentano DD, Lo Y.

A Bayesian semiparametric quantile model for longitudinal data

Xin Tong

Department of Epidemiology and Biostatistics

Arnold School of Public Health

University of South Carolina, Columbia

Quantile regression for longitudinal data assumes a parametric error distribution to get full inference according to a precious study. We propose a Bayesian semiparametric approach to random effects model using Dirichlet process mixtures for the error distribution. The flexibility of such inference under nonparametric prior models is attractive and can be incorporated to analyze longitudinal data. Mixed-effect model is a commonly used tool to account for variations across different subjects in longitudinal analysis and the Gaussian assumption of random effects distribution can be violated. Thus, it provides estimates for the subgroup specific parameters and the detection of heterogeneity in the random effects population can also be helpful as an explorative cluster analysis. Markov chain Monte Carlo sampling is used to carry out Bayesian posterior computation. Several variations of the proposed model are considered and compared via the deviance information criterion. The proposed methodology is motivated by and applied to a longitudinal growth curve data.

Nonparametric Bayesian Clustering to Detect Bipolar Methylated Genomic Loci

Xiaowei Wu

Virginia Tech

With recent development in sequencing technology, a large number of genome-wide DNA methylation studies have generated massive amounts of bisulfite sequencing data. The analysis of DNA methylation patterns helps researchers understand epigenetic regulatory mechanisms. Highly variable methylation patterns reflect stochastic fluctuations in DNA methylation, whereas well-structured methylation patterns imply deterministic methylation events. Among these methylation patterns, bipolar patterns are important as they may originate from allele-specific methylation (ASM) or cell-specific methylation (CSM).

Utilizing nonparametric Bayesian clustering followed by hypothesis testing, we have developed a novel statistical approach to identify bipolar methylated genomic regions in bisulfite sequencing data. Simulation studies demonstrate that the proposed method achieves good performance in terms of specificity and sensitivity. We apply the method to real data from mouse brain and human blood methylomes. The bipolar methylated segments detected are found highly consistent with the differentially methylated regions identified by using purified cell subsets.

Bipolar DNA methylation often indicates epigenetic heterogeneity caused by ASM or CSM. With allele-specific events filtered out or appropriately taken into account, our proposed approach sheds light on the identification of cell-specific genes/pathways under strong epigenetic control in a heterogeneous cell population.

Projection methods for analysis of imaging data of olfactory receptor neurons from the Spiny Lobster Panulirus Argus

Jun Xia[1], Manfred Schmidt [2], Charles Derby [2], Remus Osan [1, 2]

[1]Department of Mathematics and Statistics, Georgia State University

[5]Neuroscience Institute, Georgia State University

Chemoreception is an essential sense for many animals that is used for detection of food, social communication and predator avoidance. We investigate the responses properties of the Olfactory Receptor Neurons from the Spiny Lobster Panulirus Argus in response to distinct odors at different concentrations, using Ca2+ imaging recordings of populations of around 40 such neurons. We use statistical tools for comparing basic response properties (response complexity, sensitivity, selectivity) of the population of olfactory receptor neurons (ORNs) and the population of chemoreceptor neurons (CRNs) of one subsystem of 'distributed chemoreception' in Panulirus Argus. One important and advantageous aspect of characterizing response properties of chemoreceptor neurons with Ca2+ imaging is that far more records of single units are generated per time than is possible with electrophysiological methods. To extract relevant information from these imaging data in a timely manner, we automated parts our data analysis. More specifically, we have successfully implemented automatic data analysis routines in Matlab for extracting key parameters (peak amplitude of stimulus-induced Ca2+ transients; frequency and amplitude of spontaneous Ca2+ oscillations) from ΔF/F0 time-series of ORN Ca2+ recordings (change over the baseline). Furthermore, we used Principal Component Analysis (PCA) and Multiple Discriminant Analysis (MDA) to perform pattern classification of the population activity corresponding to different classes of odors at different concentrations. We find out that the population responses to different odors can be projected in a low dimensional encoding subspace where they form segregated clusters. Our future aim is to use these statistical approaches to characterize the population responses to repeated stimulation with select stimuli of different biological meaning (e.g. food stimuli vs. social communication stimuli).

Jackknife Empirical Likelihood Inference for Transformation Models

Xue Yue

Georgia State University

In the real life, transformation models are commonly used in survival analysis and economics, and maximum rank correlation estimator (MRCE) is a useful tool for dealing with this class of models. In this paper, we use jackknife empirical likelihood (JEL) method to make statistical inference for the SMRCE that is a smoothed maximum rank correlation estimator. In order to compare the JEL method and normal approximation (NA) method, simulation studies are conducted to obtain coverage probability and average length of the confidence intervals for

SMRCE. Also, in survival analysis, right censoring data is a main issue, thus we apply JEL method to smoothed maximum partial rank correlation estimator (SPRCE) as well. The simulation results show that the JEL method performs well. The proposed approach is also applied to the real data sets. Joint work with Yichuan Zhao.

Comparison of Different Computational Implementations on Fitting Generalized Linear Mixed-effects Models for Repeated Count Measures

Hui Zhang

St. Jude Children's Research Hospital

In modeling repeated count outcomes, generalized linear mixed-effects models (GLMMs) are frequently used to account for the within-cluster correlations. However, inconsistent results are usually generated by various statistical packages or procedures, especially when a moderate within-cluster correlation or overdispersion exists. We investigated the underlying numerical approaches and statistical theories that these packages or procedures are built on. Then the performances of these statistical packages and procedures were investigated by using simulated both Poisson distributed and overdispersed count data.

Computation of parameter regimes for traveling waves propagation failure in integrate and fire neural networks with periodic inhomogeneities

Jie Zhang[1], Rosahn Bhattarai [1],Xia Jun[1], Remus Osan [1, 2]
[1]Department of Mathematics and Statistics, Georgia State University
[2]Neuroscience Institute, Georgia State University

Computational models of cortical tissue often use large-scale integrate and fire neural networks to simulate the initiation and propagation of traveling waves. Often, the networks are considered in the continuous limit, with neural interactions that are homogenous and decrease exponentially as the distance between neurons increases. Here we investigate the impact of inhomogeneities on wave dynamics. For small amplitude of the inhomogeneities (ε) we determine that constant speed traveling waves are now periodically modulated, and we compute series for more of increasingly accurate approximations that are in excellent agreement with numerical

simulations. For larger amplitudes inhomogeneities, we determine analytical expressions for both the lowest ε that induces propagation failure as well as for the lowest speed of the corresponding periodically modulated activity propagation (c_ε, which can be sustained for asymptotical long times before abrupt failure), for the cases where the spatial period of inhomogeneities is either very small or very large. We used numerical simulations to confirm these results and to determine the relationship between ε and c_ε for the intermediate regimes. For future directions, we will investigate more general functions for the strong homogenous and for the periodic weak non-homogenous components of the neural interactions.

Effect of Imputation of Missing Data in the State Inpatient Databases on Racial Disparities Research

Wei Zhang[1], Andrew Gelman[2] and Stephen Lyman[3], Yan Ma[1]

(1)Department of Epidemiology and Biostatistics, Milken Institute School of Public Health, The George Washington University, Washington, DC,

(2)Columbia University, New York, NY, (3)Hospital for Special Surgery, New York, NY

Research Objective: Racial disparities (RD) in healthcare outcomes in the U.S. have been identified in recent decades for total joint arthroplasty, particularly total knee arthroplasty (TKA). We sought to study RD in TKA using the HCUP State Inpatients Databases (SID). However, as with any large scale data collection effort, the SID have a moderate amount of missing data (MD) in several patient-level variables. In particular, “patient race” has a high proportion of missingness. As a result, researchers often conduct inappropriate analysis leading to invalid inferences. This study aimed at identifying appropriate imputation methods for the SID.

Study Design: We compared five imputation methods for MD (mean imputation, random draw, hot deck, joint multiple imputation [MI], conditional MI) through a simulation constructed on real data from the SID so that the hierarchical data structures and MD patterns of the database were retained. We generated MD in a mixed types of variables including continuous (total charge), binary (sex), ordinal (household income), and unordered (race) variables. Additional predictive demographic information in patients’ places of residence and hospital characteristics were obtained from outside sources (Census, American Hospital Association) and incorporated into the imputation. To assess the performance of these methods, we reported root-mean-square error (RMSE) and bias of the imputed values with respect to the true values for continuous variables; the correctly imputed proportion for categorical variables. In addition, we formulated regression models for interesting RD outcomes including length of hospitalization (< or >=4 days) and utilization of high-volume hospitals (low, medium, high) in patients undergoing TKA. This was to assess the accuracy of coefficient estimates using imputed data.

Population Studied: Hospital discharges from Colorado in the 2005 HCUP SID.

Principal Findings: Conditional MI prediction was uniformly equivalent or superior to the best performing alternatives for all missing data structures while substantially outperforming each of the alternatives in various scenarios. Hot deck was particularly poor for binary prediction (23% correct), random draw was particularly poor for ordinal data prediction (26%), and mean imputation (5%) and joint MI (52%) did not perform well for unordered data. In addition, conditional MI had the lowest RMSEs and biases of the coefficient estimates associated with race in regression analyses, leading to the highest proportion of correct statistical significance. In contrast, mean imputation was particularly poor for assessing significance of race in utilization of high-volume hospitals (68% correct) and joint MI was particularly poor for race in length of hospitalization (32%).

Conclusions: The use of conditional MI substantially improved statistical inferences and this method outperformed other popularly used methods for MD in Colorado SID.

Founding Resource: This study was supported by the research grant (R01HS021734) funded by the Agency for Healthcare Research and Quality.

Video-Based Action Recognition Using Rate-Invariant Analysis of Covariance Trajectories

Zhengwu Zhang

Florida State University

Statistical classification of actions in videos is mostly performed by extracting relevant features, particularly covariance features, from image frames and studying time series associated with temporal evolutions of these features. A natural mathematical representation of activity videos is in form of parameterized trajectories on the covariance manifold, i.e. the set of symmetric, positive-definite matrices (SPDMs). The variable execution-rates of actions implies variable parameterizations of the resulting trajectories, and complicates their classification. Since action classes are invariant to execution rates, one requires rate-invariant metrics for comparing trajectories. A recent paper represented trajectories using their transported square-root vector fields (TSRVFs), defined by parallel translating scaled-velocity vectors of trajectories to a reference tangent space on the manifold. To avoid arbitrariness of selecting the reference and to reduce distortion introduced during this mapping, we develop a purely intrinsic approach where SPDM trajectories are represented by redefining their TSRVFs at the starting points of the trajectories, and analyzed as elements of a vector bundle on the manifold. Using a natural Riemannain metric on vector bundles of SPDMs, we compute geodesic paths and geodesic distances between trajectories in the quotient space of this vector bundle, with respect to the re-parameterization group. This makes the resulting comparison of trajectories invariant to their re-parameterization. We demonstrate this framework on two applications involving video classification: visual speech recognition or lip-reading and hand-gesture recognition. In both cases we achieve results either comparable to or better than the current literature.

Wavelet-based Imaging Genomic Modeling in Autism Spectrum Disorder Study

Hongxiao Zhu

Virginia Tech

Autism spectrum disorder (ASD) is a group of developmental disabilities that can cause significant social, communication and behavioral challenges. Although the exact causes of ASD remain unknown, research suggests that both genetic and environmental factors may be related to the development of ASD. The traditional method of diagnosis and assessment of ASD is based on verbal and physical behaviors as well as professional judgment, which can be subjective. Recent development of high-throughput genotyping and new imaging techniques make it possible to improve the assessment of ASD based on more objective information – brain activation patterns characterized by functional Magnetic Resonance Imaging (fMRI) and genetic polymorphisms. Despite the promise, integrative analysis of biomedical imaging and genetic information is difficult due to complexity of the data and lack of analytical tools. We propose a family of wavelet-based imaging genomic methods to model the three-way associations between ASD disease, fMRI images and genetic variants. The proposed methods treat the fMRI images as functional data objects and regress them on the disease variables and the genetic variables separately in a two-step procedure. In the first step, ASD-related imaging phenotypes are detected using functional data regression equipped with wavelet-space multiple testing and family-wise error rate control. These phenotypes are then fed into another functional data regression model for finding the ASD-associated genetic variants. This framework incorporates several special cases that are suitable for different data structures, including imaging data with Gaussian process random effects or random effects caused by multiple measurements per subject, images with temporal correlation, or genotype data with linkage disequilibrium. We present both frequentist and Bayesian approaches for model fitting and compare the performance of different versions of the proposed models using both simulated and real preliminary ASD data.