Program

Workshop on Biostatistics and Bioinformatics

Department of Mathematics and Statistics

Georgia State University

May 4-6, 2012

Contents

Sponsor

The Georgia State University Research Foundation, 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

Ian McKeague, Columbia University

Invited Speaker

Hua Yun Chen, University of Illinois, Chicago

Nelson Chen, Emory University

Gauri Datta, University of Georgia

Jason Fine, University of North Carolina, Chapel Hill

Leonid Hanin, Georgia Institute of Technology and Idaho State University

Eugene (Yijian) Huang, Emory University

Shan (Xianzheng) Huang, University of South Carolina

Xiaoming Huo, Georgia Institute of Technology

Zhezhen Jin, Columbia University

Cheolwoo Park, University of Georgia

Liang Peng, Georgia Institute of Technology

Jian-Jian Ren, University of Maryland - College Park

Brani Vidakovic, Georgia Institute of Technology

Lily Wang, University of Georgia

Hao Wu, Emory University

Xiangrong Yin, University of Georgia

Hongmei Zhang, University of South Carolina

Mei-Jie Zhang, Medical College of Wisconsin

Acknowledgements

The organizer thanks Earnestine Collier-Jones, Sandra Ahuama-Jonas, Yvonne Pierce , Shuman Guo, 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 4, 2012

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

Saturday, May 5, 2012

8:00-8:20 am **Registration:** on the 1st Floor, College of Education Building, 30 Pryor Street.

8:20-8:30 am Conference Welcome: Guantao Chen, Chair of the dept., Georgia State University

Opening Remarks: Mark Becker, President, Georgia State University

8:30-9:30 am **Session 1 (Keynote Talk): Chair:** *Yichuan Zhao, Georgia State University*

Analyzing Growth Trajectories

Ian McKeague, Columbia University

9:30-9:45 am **Break:** Refreshments

9:45-11:55 am **Session 2: Chair:** *Xu Zhang , Georgia State University*

Screening for Osteoporosis in Postmenopausal Women: A Case Study in Interval Censored Competing Risks Data

Jason Fine, University of North Carolina, Chapel Hill

Diagnostics of Mammograms by Wavelet-based Scaling Tools

Brani Vidakovic, Georgia Institute of Technology

Trajectory Identification with Binary Sensor Networks

Xiaoming Huo, Georgia Institute of Technology

Hierarchical Bayesian Modeling in Syndromic Surveillance

Gauri Datta, University of Georgia

11:55-1:30 pm **Lunch Time**

1:30-3:30 pm **Session 3: Chair:** *Jun Han, Georgia State University*

On Computation with the Accelerated Failure Time Model

Eugene (Yijian) Huang, Emory University

Statistical Methods for the Evaluation of Agreement and Concordance

Zhezhen Jin, Columbia University

Bivariate Nonparametric Maximum Likelihood Estimator with Right Censored Data

Jian-Jian Ren, University of Maryland - College Park

Competing Risks with Missing Covariates: Effect of Haplotypematch on Hematopoietic Cell Transplant Studies

Mei-Jie Zhang, Medical College of Wisconsin

3:30-3:45 pm **Break:** Refreshments

3:45-5:15 pm **Session 4:** **Chair:** *Yuanhui Xiao, Georgia State University*

Inference for ROC Curve and its Partial Area

Liang Peng, Georgia Institute of Technology

Practice-related Changes in Neural Activation Patterns Investigated via Wavelet-based Clustering Analysis

Cheolwoo Park, University of Georgia

Solution Paths for Large p Small n Problems

Xiangrong Yin, University of Georgia

5:15-5:30 pm Break: Refreshments

5:30-7:00 pm **Poster Session**: **Chair:** Remus Osan, Georgia State University:
Room 150, College of Education Building

7:15-9:30 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 6, 2012

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

8:30-10:30 am **Session 5:** **Chair:** *Zhezhen Jin, Columbia University*

Likelihood Inferences on Semiparametric Odds Ratio Models with Genetic Applications

Hua Yun Chen, University of Illinois, Chicago

Differential Expression in RNA-seq

Hao Wu, Emory University

Spatial Periodicities in Chromosomal Gene Expression Revealed by Spectral Analysis

Leonid Hanin, Georgia Institute of Technology

Score-based Variable Selection in Linear Measurement Error Models

Shan Huang, University of South Carolina

10:30-10:45 am **Break:** Refreshments

10:45-12:15am **Session 6:** **Chair:** *Jeff Qin, Georgia State University*

Dose Escalation with Overdose Control using a Quasi-Continuous Toxicity Score in Cancer Phase I Clinical Trials

Nelson Chen, Emory University

Nonparametric Estimation and Model Selection with Applications to Pima Indian Diabetes Study

Lily Wang, University of Georgia

Mismeasurement in Interval Estimation, Hypothesis Testing, and Variable Selection

Hongmei Zhang, University of South Carolina

Keynote Talk

Analyzing Growth Trajectories

Ian McKeague

Columbia University

Growth trajectories play a central role in life course epidemiology, often providing fundamental indicators of prenatal or childhood development, as well as an array of potential determinants of adult health outcomes. Statistical methods for the analysis of growth trajectories have been widely studied, but many challenging problems remain. Repeated measurements of length, weight and head circumference, for example, may be available on most subjects in a study, but usually only sparse temporal sampling of such variables is feasible. It can thus be challenging to gain a detailed understanding of growth velocity patterns, and smoothing techniques are inevitably needed. Moreover, the problem is exacerbated by the presence of large fluctuations in growth velocity during early infancy, and high variability between subjects. Existing approaches, however, can be inflexible due to a reliance on parametric models, and require computationally intensive methods that are unsuitable for exploratory analyses. This talk introduces a nonparametric Bayesian inversion approach to such problems, along with an R package that implements the proposed method.

Invited Talks

Likelihood Inferences on Semiparametric Odds Ratio Models with Genetic Applications

Hua Yun Chen

University of Illinois at Chicago

We propose a maximum semiparametric likelihood approach to estimation and inference on the semiparametric odds ratio model which extends the log-linear modeling framework to both discrete and continuous data. The maximum semiparametric likelihood estimator of the odds ratio parameters is shown to be consistent and asymptotically normally distributed. For the case-control type of sampling

designs with conditionally specified models, an approximate semiparametric odds ratio model is proposed to convert the difficult problem of estimating weakly identifiable parameters into a relatively easy problem of estimating strongly identifiable parameters. Statistical inference under miss-specified semiparametric odds ratio models is also established. A least square combination approach is proposed to estimate the weakly identifiable parameters when feasible, and a sensitivity analysis to the weakly identifiable parameters is proposed when the least square combination is infeasible. Simulation studies demonstrate that the proposed approach has satisfactory finite sample performance. Application of the proposed approach is illustrated by the analysis of a case-control genome-wide association study of prostate cancer.

Dose Escalation with Overdose Control using a Quasi-Continuous Toxicity Score in Cancer Phase I Clinical Trials

Nelson Chen

Emory University

In cancer Phase I clinical trials, 3+3 design is still used for its simplicity, but it has limitations such as inaccuracy of maximum tolerated dose (MTD) and inflexibility. Escalation With Overdose Control (EWOC) is a Bayesian adaptive design which can overcome these limitations and control the probability of overdosing. However, like other Phase I designs, EWOC treats toxicity response coarsely as a binary indicator (Yes vs No) of dose limiting toxicity (DLT) although patient usually has multiple toxicities and a lot of useful toxicity information is discarded. We establish a novel scoring system to treat toxicity response as a quasi-continuous variable and utilize all toxicities of patients. Our system consists of generally accepted and objective components (a logistic function, grade and type of toxicity, and whether the toxicity is DLT) so that it is relatively objective. We couple our system with EWOC to develop a new design called Escalation With Overdose Control using Normalized Equivalent Toxicity Score (EWOC-NETS) by replacing the binary indicator of DLT and the target probability of DLT with a Normalized Equivalent Toxicity Score (NETS) and a Target NETS (TNETS), respectively. Simulation studies and its application to real trial data demonstrate that EWOC-NETS can treat toxicity response as a quasi-continuous variable, fully utilize all toxicity information, and improve the accuracy of MTD and efficiency of Phase I trial. A user-friendly software of EWOC-NETS is under development and will be available in the future.

Hierarchical Bayesian Modeling in Syndromic Surveillance

Gauri Datta

University of Georgia

Syndromic surveillance is an emerging research area. Quick detection of emerging geographical clusters of disease is important to provide swift intervention to prevent a pandemic. To overcome the lag inherent in traditional public health notification structures, the Centers for Disease Control plans to implement a monitoring system to which all hospitals in USA report, in real or near-real time, admissions associated with various symptoms. Syndromic surveillance has grown opportunistically, relying on cusum charts, regression, exponentially weighted moving averages and standard time series models. A combination of time series and control charts has been used for early detection of anthrax outbreaks by tracking over-the-counter medication sales.

Reliable surveillance models are an important tool in public health because they aid in mitigating disease outbreaks, identify where and when disease outbreaks occur, and predict future occurrences. While many statistical models have been devised for surveillance purposes, none are able to simultaneously achieve important practical goals such as good sensitivity and specificity, proper use of domain information, inclusion of spatio-temporal dynamics, and transparent support to decision-makers. In an effort to achieve some of these goals, this talk proposes a spatio-temporal conditional autoregressive hidden Markov model with an absorbing state. The model performs well in both a large simulation study and in an application to influenza/pneumonia fatality data. We use a hierarchical Bayesian approach and a Poisson distribution to model the disease counts and a log-linear spatio-temporal hierarchical model to assess the posterior probability of an epidemic at each county.

Joint work with David Banks, Duke University

Matt Heaton, Duke University

Alan Karr, National Institute of Statistical Sciences

James Lynch, University of South Carolina

Francisco Vera (formerly from Clemson University)

and Frank Zou, National Institute of Statistical Sciences

Screening for Osteoporosis in Postmenopausal Women: A Case Study in Interval Censored Competing Risks Data

Jason Fine

University of North Carolina, Chapel Hill

Current US Preventive Services Task Force encourages osteoporosis screening using bone mineral density but does not specify a screening interval or ages to start and stop testing using an evidence based rationale. The current analysis explores these issues using data from the Study of Osteoporotic Fractures, the longest running cohort study of osteoporosis in the United States. Complications arise: time to osteporosis in individuals free of osteoporosis, prior fracture, and previous preventive treatment, is subject to potentially dependent censoring by fracture and preventive treatment. Endpoint definition is addressed in a competing risks framework, with a certain cumulative incidence function correctly defining the risk of osteoporosis for the screening population. The analysis of this quantity is based on intermittent bone mineral density testing. Likelihood based inference, both full and "naive", is investigated for such interval censored competing risks data, using a direct modelling strategy for the cumulative incidence functions. The screening interval is defined as a fixed time for a specified percentage of non-osteoporotic women to develop osteoporosis, accounting for the potentially dependent competing risks, which involves the use of so-called competing risks quantiles. The competing risks analysis illustrates how osteporosis risks may be precisely quantified and used to develop evidence based policy for osteoporosis screening.

Spatial Periodicities in Chromosomal Gene Expression Revealed by Spectral Analysis

Leonid Hanin

Georgia Institute of Technology and Idaho State University

The extensive effort in the analysis of gene expression over the last decade has been almost uniformly focused on the genome-wide gene expression and has largely ignored the fundamental fact that every gene has a specific chromosome location. We propose a novel method of spectral analysis for detecting hidden periodicities in gene expression signals ordered along the length of each chromosome. The method is based on a consistent estimator of the discrete spectrum of asymptotically stationary

stochastic processes previously discovered by the author for solving an entirely different problem. Using this method, we have discovered that each chromosome in rodents and humans has a unique, stable periodic pattern of gene expression insensitive to local genomic alterations. The estimated periods and amplitudes are identical for the genes located on the positive and negative DNA strands. The uncovered spatial periodicities in gene expression were found to be tissue-specific. Importantly, large differences in chromosomal gene expression in humans were observed between two normal tissues (brain and mammary gland) and their malignant counterparts (glioma and breast cancer). However, these differences are localized only on some of the chromosomes. While molecular mechanisms of chromosome-specific periodicities in gene expression have yet to be unraveled, the discovered periodic patterns of gene expression may have applications as diagnostic and prognostic tools in medicine.

On Computation with the Accelerated Failure Time Model

Eugene (Yijian) Huang

Emory University

Weighted log-rank estimating function (Tsiatis 1990) has become a standard estimation method for the censored linear regression model, or the accelerated failure time model. Well established statistically, the estimator has, however, rather poor computational properties because the estimating function is neither continuous nor, in general, monotone. Existing methods may become computationally burdensome even with moderate sample size. In this talk, we present a computationally efficient estimator through an asymptotics-guided Newton algorithm. This estimator is asymptotically equivalent to a consistent root and is barely distinguishable in practical-sized samples. But the former typically costs only thousandths to hundredths in computer time of the latter as obtained by the currently prevailing method. This tremendous improvement would facilitate broader acceptance of censored linear regression. Illustrations with clinical applications are provided.

Score-based Variable Selection in Linear Measurement Error Models

Shan (Xianzheng) Huang

University of South Carolina

We investigate variable selection procedures based on penalized score functions derived for linear regression models with error-prone covariates. Score-based tuning parameter selectors are proposed to calibrate the selection processes. It is demonstrated that the proposed variable selection procedure coupled with the new tuning parameter selector often outperforms the existing methods designed for measurement error models.

Trajectory Identification with Binary Sensor Networks

Xiaoming Huo

Georgia Institute of Technology

In office buildings, hospitals, and clinics, anonymous binary motion detectors may be installed. Trajectory identification problem is to recover object tracks, out of these binary data. We present a general approach that is based on optimization. We then consider how to incorporate false information and missing data, using maximum likelihood estimates. Robustness and reliability of the solutions will be analyzed. It has applications in building smart environments.

This is a joint work with I-Hsiang Lee (a graduate student at Georgia Tech) and Wenzhan Song (a professor in the computer science at Georgia State University).

Statistical Methods for the Evaluation of Agreement and Concordance

Zhezhen Jin

Columbia University

In medical studies, it is often of interest to assess the degree of agreement between two or more methods of clinical measurement. I will review available statistical methods for agreement assessment and discuss remaining issues and challenges. On the other hand, the concordance can be used to assess discriminative and predictive accuracy in statistical models. Recently, there has been significant development. I will review available methods and present methods for the estimation of concordance probability based on nonparametric or semiparametric modeling approach for right censored data. The asymptotic properties of the proposed estimators will be presented and the methods will be illustrated with real examples.

Practice-related Changes in Neural Activation Patterns Investigated via Wavelet-based Clustering Analysis

Cheolwoo Park

University of Georgia

The objective of this study is to evaluate brain activation using functional magnetic resonance imaging (fMRI) and activation changes across time associated with practice-related cognitive control during eye movement tasks. FMR images were acquired from participants engaged in antisaccade (generating a glance away from a cue) performance at two time points: 1) pre-test before any exposure to the task, and 2) post-test, after one week of daily practice on antisaccades, prosaccades (glancing towards a target) or fixation (maintaining gaze on a target). The three practice groups were compared across the two time points, and analyses were conducted via the application of a model-free clustering technique based on wavelet analysis. This series of procedures was developed to avoid analysis problems inherent in fMRI data and was composed of several steps: detrending, data aggregation, wavelet transform and thresholding, no trend test, principal component analysis and K-means clustering. The main clustering algorithm was built in the wavelet domain to account for temporal correlation. We applied a no trend test based on wavelets to significantly reduce the high dimension of the data. We clustered the thresholded wavelet coefficients of the remaining voxels using the principal component analysis K-means clustering. Over the series of analyses, we found that the antisaccade practice group was the only group to show decreased activation from pre- to post-test in saccadic circuitry, particularly evident in supplementary eye field, frontal eye field, precuneus, and cuneus.

Inference for ROC Curve and its Partial Area

Liang Peng

Georgia Institute of Technology

First we give a correct formula for the asymptotic variance of the nonparametric estimate of the partial area under the receiver operating characteristic curve, and then provide a consistent estimate for the asymptotic variance. Second we propose a smooth jackknife empirical likelihood method to construct a confidence interval for the ROC curve without estimating any additional quantities.

Bivariate Nonparametric Maximum Likelihood Estimator with Right Censored Data

Jian-Jian Ren

University of Maryland - College Park

In the analysis of survival data, we often encounter situations where the response variable (the survival time) T is subject to right censoring, but the covariates Z are completely observable. To use the nonparametric approach (i.e., without imposing any model assumptions) in the study of the relation between the right censored response variable T and the completely observable covariate variable Z, one natural thing to do is to estimate the bivariate distribution function F_o(t, z) of (T, Z) based on the available bivariate data which are right censored in one coordinate - we called it BD1RC data. In this article, we derive the bivariate nonparametric maximum likelihood estimator (BNPMLE) F_n(t,z) for F_o(t, z) based on the BD1RC data, which has an explicit expression and is unique in the sense of empirical likelihood. Other nice features of F_n(t,z) include that it has only nonnegative probability masses, thus it is monotone in bivariate sense, while these properties generally do not hold for most existing distribution estimators with censored bivariate data. We show that under BNPMLE F_n(t,z), the conditional distribution function (d.f.) of T given Z is of the same form as the Kaplan-Meier estimator for

the univariate case, and that the marginal d.f. F_n(\infty,z) coincides with the empirical d.f. of the covariate sample. We also show that when there is no censoring, F_n(t,z) coincides with the bivariate empirical distribution function. For the case with discrete covariate Z, the strong consistency and weak convergence of F_n(t,z) are established. The extension of our BNPMLE F_n(t,z) to the case with p-variate Z for p>1 is straightforward.

This work is joint with Tonya Riddlesworth.

Diagnostics of Mammograms by Wavelet-based Scaling Tools

Brani Vidakovic

Georgia Institute of Technology and Emory University

We present results from a comparative investigation into the diagnostic performance of several wavelet-based estimators of scaling, some from published literature and some newly proposed. 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. We found that Abry-Veitch and modified weighted Theil-type estimators provided the best classification rates, while the standard wavelet-based OLS estimator performed worst. The results are robust with respect to choice of wavelets (Haar wavelet being an exception) and are of potential clinical value. The diagnostic is based on the properties of image backgrounds (which is an unused diagnostic modality in mammograms) and the best average correct classification rates achieve 90%, varying slightly with the choice of basis, levels used, and size of training set.

This work is joint with Erin Hamilton and Seonghye Jeon (GaTech), Kichun Lee (Hanyang University, Korea) and Pepa Ramirez (University of Sevile, Spain).

Nonparametric Estimation and Model Selection with Applications to Pima Indian Diabetes Study

Lily Wang

University of Georgia

Arizona's Pima Indians have the world's highest rate of diabetes, and the rest of the world is catching up fast. Why do so many Pima Indians have diabetes? We explore the impact of several variables on the possibility of having a positive test. Regression-based methods assuming a linear relationship between the covariates and the probability of a positive test is the commonly used approach. However, a closer investigation shows that the effect of AGE and BMI on the logit transformation of the probability of a positive test may be nonlinear. We study a generalized additive partial linear model, proposing the use of polynomial spline smoothing for estimation of nonparametric functions, and deriving quasi-likelihood based estimators for the linear parameters. We develop a class of variable selection procedures to identify significant factors, which is shown to have an oracle property.

Differential Expression in RNA-seq

Hao Wu

Emory University

Recent developments in RNA-sequencing (RNA-seq) technology have led to a rapid increase in gene expression data in the form of counts. RNA-seq can be used for a variety of applications, however, identifying differential expression (DE) remains a key task in functional genomics. There have been a number of statistical methods for DE detection for RNA-seq data. One common feature of several leading methods is the negative binomial (gamma-Poisson mixture) model. The distinct feature in various methods is how the variance, or dispersion, in the gamma distribution is modeled and estimated. We evaluated several large public RNA-seq datasets and find that the estimated dispersion in existing methods does not adequately capture the heterogeneity of biological variance among samples. We present a new empirical Bayes shrinkage estimate of the dispersion parameters and demonstrate improved DE detection.

Solution Paths for Large p Small n Problems

Xiangrong Yin

University of Georgia

In this talk, I will introduce a new but very simple framework to solve the large p small n problems. The framework decomposes the data into pieces so that existing methods can be applied. We propose two separate paths to implement the framework. Our paths provide sufficient procedures for identifying informative variables sequentially. The paths are very general and we shall illustrate their efficacy via simulation and a real data set by using sufficient dimension reduction and variable selection methods.

Joint work with Haileab Hilafu

Mismeasurement in Interval Estimation, Hypothesis Testing, and Variable Selection

Hongmei Zhang

University of South Carolina

Mismeasurement includes misclassification on categorical variables and measurement error in continuous variables. In some situations, only false-positive errors could occur, which made the correction of misclassification possible when inferring population proportions. Measurement errors, on the other hand, can cause power loss in hypothesis testing and false selection of important variables. In this talk, under the Bayesian framework, we will discuss methods dealing with mismeasurement in different situations including interval estimation, hypothesis testing and variable selection.

Competing Risks with Missing Covariates: Effect of Haplotypematch on Hematopoietic Cell Transplant Studies

Mei-Jie Zhang

Medical College of Wisconsin

In this talk we consider a problem from hematopoietic cell transplant (HCT) studies where there is interest on assessing the effect of haplotype match for donor and patient on the cumulative incidence function for a right censored competing risks data. For the HCT study, donor's and patient's genotype are fully observed and their haplotypes are missing. In this talk we describe how to deal with missing covariates of each individual for competing risks data. We suggest a procedure for estimating the cumulative incidence functions for a flexible class of regression models when there are missing data, and establish the large sample properties. The proposed approach has been applied to a HCT example.

Poster Abstracts

A Comparison of Three Approaches for Constructing Robust Experimental Designs

Vincent Agboto

Meharry Medical College

While optimal designs are commonly used in the design of experiments, the optimality of those designs frequently depends on the form of an assumed model. Several useful criteria have been proposed to reduce such dependence, and efficient designs have been then constructed based on the criteria, often algorithmically. In the model robust design paradigm, a space of possible models is specified and designs are sought that are efficient for all models in the space. The Bayesian criterion given by DuMouchel and Jones (1994), posits a single model that contains both primary and potential terms.

In this article we propose a new Bayesian model robustness criterion that combines aspects of both of these approaches. We then evaluate the efficacy of these three alternatives empirically. We conclude that the model robust criteria generally lead to improved robustness; however, the increased robustness can come at a significant cost in terms of computing requirements.

An Integer Programming Approach to Novel Transcript Reconstruction from Paired-End RNA-Seq Reads

Adrian Caciula

Georgia State University

In this work, we propose a novel statistical “genome guided” method called “Transcriptome Reconstruction using Integer Programming” (TRIP) that incorporates fragment length distribution into

novel transcript reconstruction from paired-end RNA-Seq reads. To reconstruct novel transcripts, we create a splice graph based an exact annotation of exon boundaries and RNA-Seq reads. The exact annotation of exons can be obtained from annotation databases (e.g., Ensembl) or can be inferred from aligned RNA-Seq reads. A splice graph is a directed acyclic graph (DAG), whose vertices represent exons and edges represent splicing events. We enumerate all maximal paths in the splice graph using a depth-first-search (DFS) algorithm. These paths correspond to putative transcripts and are the input for the TRIP algorithm.

A Novel Phase II Design to Minimize Trial Duration and Improve the Success Rate of Follow-up Phase III Trial

Ye Cui

Georgia State University

A Phase II trial is an expeditious and low cost trial with the primary goal of screening potentially effective agents prior to confirmatory Phase III trial. The success rate of Phase III oncology trials remains very low despite the success demonstrated in the preceding Phase II trials. This discordance is mainly due to the different endpoints used in Phase II (disease response) and III (survival) trials. While a robust disease response is expected to translate into survival improvement, this is NOT guaranteed. Moreover, disease response can be determined quickly whereas survival estimation requires a long period of follow up. We propose a novel 2-stage screening design for phase II trials whereby disease response endpoint is used as an initial screening to select potentially effective agents within a short time interval followed by a second screening stage where progression free and/or overall survival is estimated to confirm the efficacy of agents. This design can improve trial efficiency and reduce cost by stopping the evaluation of an ineffective agent based on low response rate. The second survival endpoint screening will substantially increase the success rate of follow-up Phase III trial by using the same outcomes. We conducted simulation studies to investigate the underlying statistical considerations to optimize the significant levels of the two screening stages in the design. A real Phase II clinical trial data was reanalyzed for real world assessment of the design characteristics.

Integrative, Multi-Platform, Whole-Genome Array Analyses Refine the Molecular Signature of Multiple Myeloma

Khanjan Gandhi

Emory University

Recent advances in genomics have led to a plethora of new technologies, platforms and analysis methodologies. As a result, it can become quite confusing to identify the most appropriate technology to use to address a specific biological question. To facilitate this process for researchers, we are developing a workflow for analysis of data generated by five different, vastly popular platforms (microarray gene expression, methylation, miRNA, SNP copy number, RNAseq) that is supplemented by the use of public repository data. Our workflow integrates data from various array types to address the question of potential mechanisms underlying observed changes in gene expression and as such, may be considered a complimentary approach to next generation sequence data though not without cost in terms of several needed array types, analytical development for their integration and interpretation, and synthesizing results. This work was motivated by a study of Multiple Myeloma (MM) that used three cell lines (KMS11, MM1s, RPMI8226) applied to all five platforms. Based on a defined list of genes whose expression is specific to each cell line and common among cell lines, we investigate potential mechanisms underlying observed cell line specific expression changes through examining cell line specific changes in the other available array types. Results in terms of expression and methylation changes are able to be validated by the use of more dense arrays and external, publicly available data. Integrated analysis on multiple platforms will help to understand the biological mechanisms behind the observed differences between these three model systems of MM.

Jackknife Empirical Likelihood Confidence Intervals of the Gini Index

Dirk Gilmore

Georgia State University

The Gini index is one the most popular measures of income inequality across individuals within a population or across populations. Previously, empirical likelihood methods have proven an effective

tool to produce confidence intervals of this measure. The resultant limiting distribution, however, is a scaled chi –squared distribution. In order to improve the efficiency, this study offers modifications of existing methods of generating these confidence intervals. Specifically, two jackknife empirical likelihood (JEL) methods are employed. Using simulated data, both JEL methods result in improved average lengths (AL) of the confidence intervals with one method (JEL2) providing slightly better coverage probabilities. These methods are also shown to perform well using real data.

Sufficient Dimension Reduction and Statistical Modeling of Plasma Concentrations

Haileab Hilafu

University of Georgia

Many dietary and biochemical epidemiologic studies have shown an inverse association between beta-carotene in human plasma and the risk of cancer. Thus it is crucial to find factors influencing beta-carotene levels in human plasma. Personal characteristics and dietary factors are easy to be collected, and then it is important to build models using these variables to predict and evaluate plasma concentrations of retinol and beta-carotene accurately. Motivated by this, we develop a novel dimension reduction method, together with variable selection, to deal with a scenario where data has multivariate responses, continuous predictors and categorical predictor variables. Simulations show the efficacy of our approach. We apply our method to a data from Nierenberg et al. (1989) and show different structures from previous analysis of the data.

Experimental Phase Relation Captured by Model Central Pattern Generator

Sajiya Jalil (presenting), Dane Allen, Andrey Shilnikov

Georgia State University

Central pattern generators (CPGs), that are groups of neurons forming small networks via synaptic connections, can be identified by observing their activity patterns in behaving animals. In our study, we explore a plausible network that represent swim CPG in the marine invertebrate – Melibe leonina. We employ mathematical models developed using Hodgkin-Huxley formalism with parameter estimation from leech heart interneurons. The model of leech interneurons has been studied extensively and shown, both mathematically and experimentally, to have the ability to transition into a number of distinct patterns including square wave bursting, spiking, and chaos. In addition, multistability with two or more coexisting stable patterns has been reported for this model. We design the CPG models inspired by the specific phase relations seen in the experimental voltage traces. We include four neurons, connected via fast non-delayed inhibitory synapses modeled by fast threshold modulating function (FTM). Due to intrinsic symmetry, the network can be treated as two pairs of half center oscillators (HCOs). In the HCO, neurons reciprocally inhibit each other, leading to activity patterns that alternate. We consider unidirectional inhibition between the pairs of HCOs, and find phase-locked state that is idiosyncratic of the experimental system. We identify control parameters for the pattern in question, which corresponds to a single attractor for the phase-lag return mapping on a 3D torus. Our goal is to explain the mechanism that causes the particular phase-locked state and explore parametric regime for sensitivity and emergence of additional patterns in the system. In the future, we plan to enhance the CPG models by including extra interneurons ans synapses of other types, introducing heterogeneity in network connections and by increasing physiological fine details that are currently neglected. Mechanistic understanding of CPGs is important for engineering equipments that are dynamically controlled by circuits, such as in robotics and prosthetics.

Empirical Likelihood Inference for the Bivariate Survival Function under Univariate Censoring

Ali Jinnah

Georgia State University

Lin and Ying (1993) proposed a non-parametric estimator of the bivariate survival function of paired failure times under univariate censoring. Recently Lu and Burke (2008) proposed estimators for bivariate distribution of paired failure times under univariate censoring. In this paper, we apply an empirical likelihood by estimating the survival function of censored time by Kaplan-Meier estimator. Adjusted empirical likelihood (AEL) confidence intervals for the bivariate survival function are obtained. Furthermore, we conduct a simulation study to evaluate the performance of the proposed AEL method and the normal approximation (NA) method in terms of coverage probability and interval length. The simulation study shows the coverage probabilities using the AEL method obtained are close to the

nominal levels. We compare the results with the probability coverage and interval lengths. Coverage probability obtained by Lin and Ying (1993) is better .

Joint work with Yichuan Zhao

Multi-Commodity Flow Methods for Quasispecies Spectrum Reconstruction Given Amplicon Reads

Nicholas Mancuso

Georgia State University

RNA viruses depend on error-prone transcriptase for replication within an infected host. These errors lead to a high mutation rate which creates a diverse population of closely related variants [1]. This viral population is known as a quasispecies. As breakthroughs in nextgeneration sequencing have allowed for researchers to apply sequencing to new areas, studying genomes of viral quasispecies is now realizable. By understanding the quasispecies, more effective drugs and vaccines can be manufactured as well as cost- saving metrics for infected patients implemented [2]. Reconstructing the quasispecies spectrum is difficult for several reasons. The actual amount of variants may be obfuscated by conserved regions in the genome that extend beyond the maximum read length. Additionally, the amount of possible assignments of reads to variants in overlapping segments grows quickly. Furthermore, we are required to rank the variants by frequency. Previous approaches have utilized min-cost flows, probabilistic methods, shortest paths, and population diversity for the quasispecies spectrum assembly problem [3], [6], [5], [4].

We solve the problem of quasispecies spectrum reconstruction by utilizing multi-commodity unsplittable flows. The problem space is formulated as a read graph similarly to [7] and then the minimum amount of unsplittable flows covering all reads is found using Integer Programming.

References:

[1] Duarte EA, Novella IS, Weaver SC, Domingo E, Wain-Hobson S, Clarke DK, Moya A, Elena SF,

de la Torre JC, Holland JJ. (1994), RNA virus quasispecies:significance for viral disease and

epidemiology.

[2] Skums Pavel, Dimitrova Zoya, Campo David S., Vaughan Gilberto, Rossi Livia, Forbi Joseph

C, Yokosawa Jonny, Zelikovsky Alex, Khudyakov Yury. (2011). Efficient error correction for nextgeneration

sequencing of viral amplicons. International Symposium on Bioinformatics Research

and Applications

[3] Westbrooks K, Astrovskaya I, Campo D, Khudyakov Y, Berman P, Zelikovsky A: HCV

quasispecies assembly using network flows. In Proc. ISBRA 2008:159-170.

[4] Prosperi MC, Prosperi L, Bruselles A, Abbate I, Rozera G, Vincenti D, Solmone MC,

Capobianchi MR, Ulivi G: Combinatorial analysis and algorithms for quasispecies reconstruction

using next-generation sequencing. BMC Bioin-formatics(2011), 12(1):5.

[5] Astrovskaya I., Tork, B., Mangul, S., Westbrooks, K.,

M˙andoiu, I., Balfe, P., Zelikovsky A (2011) Inferring Viral Quasispecies Spectra from 454

Pyrosequencing Reads. BMC Bioinformatics 12

6] Zagordi O, Klein R, Daumer M, Beerenwinkel N: Error correction of next-generation

sequencing data and reliable estimation of HIV quasispecies. Nucleic Acids Research(2010),

38(21):7400-7409

[7] Mancuso, N., Tork, B., Skums, P., Mandoiu, I., Zelikovsky, A. Viral Quasispecies

Reconstruction from Amplicon 454 Pyrosequencing Reads. Bioinformatics and Biomedicine

Workshops (BIBMW), 2011 IEEE International Conference on. (2011)

Improve Medical Diagnosis with Multi-class Ramp-Loss SVM

Piyaphol Phoungphol (presenting) , Yanqing Zhang, and Yichuan Zhao

Georgia State University

Medical Diagnosis is an expensive, time-consuming and really subjective process due to different medical experts' opinions. Using machine learning algorithm to predict medical conditions can help substantially reduce healthcare costs and increase the accuracy of diagnosis.

In our research, we use our new multi-class SVM with ramp loss to predict heart disease in patients. Our new algorithm utilizes a new loss function called Ramp-loss function, combined with Weighted cost method to handle imbalance and noise in medical diagnosis data. The experiment result shows that our new approach dramatically improves prediction accuracy over traditional multiclass SVM.

Genome Scaffolding via Integer Linear Programming and SPQR-Tree Decomposition

Hamed Salooti

Georgia State University

Since the completion of the human genome, rapid advances in high-throughput sequencing (HTS) technologies have resulted in orders of magnitude higher throughput and lower cost compared to classic Sanger sequencing . Indeed, top-of-the-line instruments from Illumina and Life Technologies are currently able to generate in a single run billions of reads with a total length of hundreds of gigabases, at a cost of mere cents per megabase. Short contig lengths, typically between 2-4Kb, are also the characteristic of tens of draft genome assemblies generated from low-coverage (2x) Sanger, (<10x HTS) reads over the past decade. To increase the utility of such fragmented assemblies, additional long-range linkage information is used to orient contigs relative to one another and order them in larger structures referred to as scaffolds. Unfortunately linkage information provided by HTS pairs is noisy due to both chimeric pairs resulting from library preparation artifacts and erroneous mapping of reads originating from repeats. These difficulties, along with the sheer number of HTS pairs and contigs that must be handled, render scaffolding methods developed for Sanger pairs ineffective on HTS data.

Determining the orientation of contigs is an NP-hard problem . Therefore, feasible methods to solve this problem are restricted to heuristics resulting approximation solutions. We present a scaffolding algorithm with its quality assessed using a draft assembly of the Venter genome. An integer linear

program (ILP) is developed to handle the scaffolding problem which involves using long range linkage information obtained from mapped mate pairs to find orientations and linear orderings that maximize a likelihood function. This algorithm is scalable since it exploits the sparsity of the resulting scaffolding graph. Utilization of non-serial dynamical programming (NSDP) allows us to compute solutions for small components (coming from SPQR-Tree decomposition) and reconstruct the optimal solution.

Simulation Support for Principled Sure Independence Screening of Ultra-high Dimensional Covariates for the Additive Risk Model

Jay Schamel

Georgia State University

A critical area of study in covariate survival analysis is model reduction when the number of parameters far exceeds the available data. In the clinical setting, this situation often arises in microarray studies to identify genetic effects on survival data. In the last decade, many methods for high-dimensional variable selection have been adapted from generalized regression to censored survival analysis, including penalized regression and marginal screening. Zhao and Li (2010) developed a principled approach to marginal sure independence screening for the Cox model (PSIS). In this study-in-progress, we use simulation to explore the possibility of extending their method to the additive risk model of Lin and Ying. We also analyze data from a multiple myeloma treatment study, and compare the results with those obtained by Zhao and Li. In both simulated and real-world data, we find that the additive risk version of PSIS retains the desirable properties found with the Cox model: the PSIS tuning parameter can be interpreted as the desired false positive rate, pre-screening with PSIS can be used to increase the performance and accuracy of penalized regression selection methods, and the method is robust to model misspecification. However, under the additive risk model, PSIS appears to have higher false negative rates than under the Cox model. Currently, this issue is being explored.

Direction Estimation in Single-index Models via Distance Covariance

Wenhui Sheng

University of Georgia

We propose a new method for estimating the direction in the single-index model based on distance covariance. This method needs no parametric or semi-parametric assumptions. It does not require the existence of density (jointly, marginally or conditionally); and no smoothness is needed in the estimation procedure. Furthermore, the method efficiently estimates the direction in single-index model with continuous predictors alone. Under regularity conditions, we show that our estimator is root-n consistent and asymptotically normal. We compare the performance of our method with those of existing dimension reduction methods by simulation and find strong evidence of its advantage over a wide range of models. The method consistently achieves higher accuracy in simulations we consider. Finally, we apply our method to analyze a real data set concerning air pollution.

Sparse Optimal Discriminant Clustering

Yanhong Wang

Georgia State University

Optimal discriminant clustering (ODC) is a clustering method using the optimal scoring and the ridge penalty. It performs well for low-dimensional data, however, for high-dimensional data, usually many of the features are non-informative and including all of them makes ODC unsuccessful. Here we propose the sparse optimal discriminant clustering (SODC) by adding a type of group LASSO penalty on the original ODC. To perform the method, we develop an efficient block coordinate descent algorithm, along with a method for selecting tuning parameters based on clustering stability. The effectiveness of the method is examined through numerical simulations and two real applications.

Joint work with Yixin Fang, New York University and Junhui Wang, University of Illinois at Chicago

Advanced Designs for Phase I Cancer Clinical Trials

Zhibo Wang

Emory University

A Phase I clinical trial is the first trial for a new agent or investigational treatment on humans with the primary objective to estimate a Maximum Tolerated Dose (MTD) that maximizes therapeutic effect under the highest acceptable toxicity level that is safe. The choice of an efficient and optimal Phase I design presents a challenging problem. Standard 3+3 design remains the most widely used because of its simplicity, robustness, and easy implementation, despite its inaccurate MTD and inability to estimate an MTD with probability of dose limiting toxicity (DLT) <20% or >33%. Escalation with Overdose Control (EWOC), Continuous Reassessment Method (CRM), and Isotonic Design (ID) are modern adaptive designs with flexibility, greater MTD accuracy, and improved trial efficiency. In our research, we further extend these novel Phase I designs to address the following upcoming needs for contemporary Phase I clinical trials: 1) improvement in MTD accuracy and trial efficiency by fully utilizing all patient toxicities by use of our novel Normalized Equivalent Toxicity Scoring (NETS); 2) estimation of a personalized MTD, a first and critical step towards personalized medicine, by adjusting for patients' characteristics and genomic profile; 3) utilization of time to toxicity information in dose allocation without waiting period by using a weight to account for the fraction of toxicity monitoring time when the quasi-continuous toxicity score of patient is measured. To facilitate the implementation of Phase I trials using our new designs, we are developing corresponding user friendly software available as freeware from our website.

Stability Selection in Dimension Reduction

Wenbo Wu

University of Georgia

Many existing methods, such as sparse eigen-decomposition estimation (SED) and sparse sufficient dimension reduction (SSDR) transformed a dimension reduction problem into a regression-type formulation and adopted shrinkage estimation procedures to produce sparse and accurate solutions. But such shrinkage estimations can sometimes be inconsistent without choosing the effective tuning parameters. We propose a stability selection method which is based on sub-sampling in combination with random weights selection to widen the effective tuning parameter range which will hence lead to a more consistent estimation procedure. Stability selection can be directly incorporated with SED and SSDR procedure to provide more accurate and more consistent estimation of dimension and directions in central subspace. Simulation results demonstrate the effectiveness of stability selection that used in dimension reduction.

On Second Thought: Uncertainty, Learning and Lapsing Behavior in Long-term Care Insurance Market

Andinet Woldemichael

Georgia State University

Many log-term care insurance policy holders voluntarily drop their policy before filing a claim. The reasons behind this behavior are not adequately explored. This study empirically investigates if initial uncertainty about own risk type and ex-post learning affects subsequent lapse decision. The main premise is that some individuals initially perceive themselves to be of higher (lower) risk type than they truly are and theory suggest that those who learn that they are of lower risk type than anticipated tend to drop their insurance plan sooner than later. Using panel data from the Health and Retirement Study (HRS), I implement empirical models which accommodate the learning process and estimate the probability and timing of lapse. These behaviorally rich models are estimated using Markov Chain Monte Carol (MCMC) methods in a Bayesian framework. The result suggests that, indeed, individuals who ex-post believe to have lower risk of disability than initially anticipated tend to lapse their policy sooner than later. Also, those who had lower initial degree of confidence about their risk type have higher probability of lapsing. However, some individuals tend to keep their policy even if they believe that they have lower risk of disability than initially anticipated suggesting heterogeneity in risk perception and preference.

Model-based Method For Analyzing HGS Data

Meng Zhao

Emory University

The next generation sequencing (NGS) technologies have been rapidly adopted in an array of diverse applications. Although extremely promising, the massive amount of data generated from NGS, substantial biases and correlation pose daunting challenges for data analysis. By treating observed data as random samples from probability distributions, model-based methods can accommodate uncertainties explicitly and also automatically leads to rigorous statistical inference. Inspired by the success of model-based methods in the analysis of other high throughput genomics data such as microarray, we strived to develop novel model-based methods to analyze the complicated data generated from the new NGS-based experiments, aiming to help biologists to extract new biological insights from massive NGS data.