Second Annual Twin Cities ASA

FALL RESEARCH CONFERENCE

MEDTRONIC, MOUNDS VIEW, MN

Monday, October 3, 2016

Graduate Student Posters

An adaptive association test for microbiome data

Chong Wu, Division of Biostatistics University of Minnesota

There is increasing interest in investigating how the compositions of microbial communities are associated with human health and disease. Although existing methods have identified many associations, a proper choice of a phylogenetic distance is critical for the power of these methods. To assess an overall association between the composition of a microbial community and an outcome of interest, we present a novel multivariate testing method called aMiSPU, that is joint and highly adaptive over all observed taxa and thus high powered across various scenarios, alleviating the issue with the choice of a phylogenetic distance. Our simulations and real-data analyses demonstrated that the aMiSPU test was often more powerful than several competing methods while correctly controlling type I error rates. The R package MiSPU is available at https://github.com/ChongWu-Biostat/MiSPU and CRAN.

Induced Smoothing for Rank-based Regression with Recurrent Gap Time Data

Tianmeng Lyu, Division of Biostatistics, University of Minnesota

A broad class of semiparametric regression models has recently been proposed for the analysis of gap times between consecutive recurrent events. Among them, the accelerated failure time (AFT) model is especially appealing to statistical practitioners owing to its direct interpretation of the covariate effects. However, the rank-based estimating function for the AFT model is a non-smooth step function of regression parameters, and hence, the corresponding estimators may not be well defined. Moreover, the popular resampling or perturbation based variance estimation for the AFT model requires solving rank-based estimating equations repeatedly and hence can be computationally challenging. In this paper, we extend the induced smoothing method to the AFT model for recurrent gap time data and propose a smooth estimating function, which permits the application of standard numerical methods. Large sample properties and an efficient variance estimator are provided for the proposed smooth estimating function method. We also propose to apply computationally efficient methods to improve the variance estimation for the estimators from the non-smooth rank-based estimation equations. Simulation studies and a data application are presented to compare the performance of the non-smooth and the proposed smooth estimating functions with various variance estimation methods.

Investigating Efficacy of Learning Procedures for Children with Primary Language Impairment

Sakshi Arya, School of Statistics, University of Minnesota

This poster is based on my summer consulting project with the Department of Speech, Language and Hearing Sciences. Current grammatical treatment approaches for children with primary language Impairment (PLI) yield only moderately significant gains after extensive treatment periods and are thus, inadequate. One of the core language weaknesses of children with PLI is poor use of grammatical forms with such weaknesses persisting well into adolescence. Traditional treatments use inductive approaches (e.g., providing models of problematic forms at a high frequency) in which the learner is expected to implicitly acquire and generalize target grammatical forms. Unlike traditional inductive approaches, deductive instruction aims to make the learner explicitly aware of the underlying language pattern by directly presenting the pattern or pedagogic rule. The goal of the proposed study is to compare the efficacy of a traditional inductive approach to an approach that incorporates deductive instruction. Guided by strong preliminary data, this data will be tested through four specific aims: (a) to determine if a deductive approach is more efficacious than the inductive approach when teaching novel grammatical forms; (b) to determine the longitudinal effects of the two treatments and compare them based on efficiency; (c) to determine whether a deductive approach is differentially effective when teaching three novel grammatical forms; and (d) to determine if a deductive approach is differentially effective at teaching children with PLI who have different level of learning disability based on some comprehensive tests done before the study and also based on other covariates. We use longitudinal mixed effect models to answer most of these questions.

Reproducibility of High-Dimensional Variable Selection Method

Wenjing Yang, School of Statistics, University of Minnesota

The use of data mining methods to identify important and useful variables under a candidate set for a response variable is now prevalent in bioinformatics and many other fields. Previous publications have already introduced various model selection methods such as LASSO, MCP, and SCAD in dealing with high-dimensional data and have demonstrated helpful predictive performances. However, with exponentially large number of predictors and limited sample size, the selected model can be highly unstable. Thus, the evidence of making strong arguments about the reliability and consistency of the identified selection of variables is weak and unreliable. This research brings multiple variable selection methods into simultaneous consideration and uses variable selection deviation measure for evaluation to obtain a new variable selection approach that produces effective, reproducible results.

Statistics in Psychology: studying trauma and stress

Haema Nilakanta, School of Statistics, University of Minnesota

This poster is based on my summer consulting experience with the University of Minnesota, Twin Cities (UMN) School of Statistics' Consulting Center and the Department of Psychology.

Counseling psychologists are interested in understanding the effects of traumatic events, stress management and gender on emotional and distress levels of individuals. During the Summer of 2016 I worked with the Frazier Lab, Dr. Patricia Frazier's lab in Counseling Psychology, on two projects investigating this relationship. The two projects are briefly described below:

1) Daily Diary Study: focused on the analyses of longitudinal survey data collected from 268 students in an Introduction to Psychology class over a two-week period. The goal of the study was to investigate how trauma history and sex of a student affects their emotional and distress levels on a daily basis and over the course of the study. Our preliminary results show that some traumas, such as emotional abuse, have a positive relationship with the experience of certain daily stressors, meanwhile others traumas, such as bereavement do not. We also generally saw that females experience more stressors than males. In addition, females have a greater odds of reporting interpersonal stress (stress related to family and relationship problems) than males.

2) Combined Intervention Study: a meta-analysis of longitudinal data from multiple studies investigating online stress management interventions for college students. We explored fitting latent growth mixture models to classify the intervention student trajectories. While still exploring these models, there seems to be an indication that when fitting three latent classes, there exists one class which contains a small subset of students who have an ideal drop in stress level over time.

This poster also serves to illustrate the overall structure and experience of my summer consulting experience.

Credible Subgroup Inference for Bounding the Benefiting Subpopulation for Many Treatments and Multiple Endpoints

Patrick Schnell, Division of Biostatistics, University of Minnesota

Many new experimental treatments outperform the current standard only on a subset of the population. The credible subgroups method provides a pair of bounding subgroups for the benefiting subgroup constructed so that one contains only patients with an expected benefit and the other contains all patients with an expected benefit. However, when more than two treatments and multiple endpoints are under consideration, there are many possible requirements for a particular treatment to be beneficial. We extend the credible subgroups method to handle such cases using a concept derived from admissibility, and apply the extended method to an example dataset from an Alzheimer's disease treatment trial.

Generalizing the Matrix Normal Distribution An application to spatio-temporal data

Karl Oskar Ekvall, Brian Gray, School of Statistics, University of Minnesota

We consider a generalization of the matrix normal distribution for the error vector in multivariate linear models. Our generalization assumes that the correlation structure is separable in the sense of a Kronecker product but, in contrast to the matrix normal, allows full variance heterogeneity. We propose a blockwise coordinate descent algorithm for maximizing the likelihood under our model. The work is motivated by an application to spatio-temporal data comprised of 20 years of quarterly water temperature measurements at

20 locations on the Mississippi river. The interest is in modeling temperature trends, taking into account both spatial and temporal dependence.

ThrEEBoost: Thresholded Boosting for Variable Selection and Prediction via Estimating Equations

Ben Brown, Division of Biostatistics, University of Minnesota

Most variable selection techniques for high-dimensional models are designed to be used in settings where observations are independent and completely observed. At the same time, there is a rich literature on approaches to estimation of low-dimensional parameters in the presence of correlation, missingness, measurement error, selection bias, and other characteristics of real data. In this paper, we present ThrEEBoost (Thresholded EEBoost), a general-purpose variable selection technique which can accommodate such problem characteristics by replacing the gradient of the loss by an estimating function. ThrEEBoost generalizes the previously-proposed EEBoost algorithm (Wolfson, 2011) by allowing the number of regression coefficients updated at each step to be controlled by a thresholding parameter. Different thresholding parameter values yield different variable selection paths, greatly diversifying the set of models that can be explored; the optimal degree of thresholding can be chosen by cross-validation. ThrEEBoost was evaluated using simulation studies to assess the effects of different threshold values on prediction error, sensitivity, specificity, and the number of iterations to identify minimum prediction error under both sparse and non-sparse true models with correlated continuous outcomes. We show that when the true model is sparse, ThrEEBoost achieves similar prediction error to EEBoost while requiring fewer iterations to locate the set of coefficients yielding the minimum error. When the true model is less sparse, ThrEEBoost has lower prediction error than EEBoost and also finds the point yielding the minimum error more quickly. The technique is illustrated by applying it to the problem of identifying predictors of weight change in a longitudinal nutrition study.

Investigating Carp Movement Pattern with Statistics

Yannan Pan, Xiaowan Liu, School of Statistics, University of Minnesota

Carp, an inland freshwater fish first introduced to North America for food purpose, has become an infamous aquatic invasive species and did damage to the aquatic ecosystems of Minnesota. To gain understanding and control over carp, the researchers from Fisheries Department collaborated with us to discover carp's moving patterns. Our goal is to discover if there is a changing moving pattern on a 24-hour scale, affected by factors such as phase, day and night difference, sex, lake region and individual differences.

The experiment designed by our clients is a longitudinal study where the response distance data are measured repeatedly over period (6 periods a day); considering we also need to treat individual variation as a random effect, linear mixed effects model is used. The statistical analysis has shown us some interesting results. The carp's activity is significantly related with phase and lake. Overall, carp are most active during migration phase (April) in shallow, and less active during aggregation phase (Jan-March) in long lake; carp are more active at night during aggregation phase in long lake; males are more active during spawning phase (May-June) in shallow area.

Adaptive testing of SNP-brain functional connectivity association via a modular network analysis

Chen Gao, Junghi Kim and Wei Pan Division of Biostatistics, University of Minnesota

Although several methods exist for estimating brain functional networks, such as the sample corre- lation matrix or graphical lasso for a sparse precision matrix, they may not yield network estimates with scale-free topology and/or network modularity, which have been demonstrated to be present in brain function networks. In particular, alteration of brain modularity is observed in patients suffering from various types of brain malfunctions. We adapt a weighted gene co-expression network analysis (WGCNA) framework to resting-state fMRI (rs-fMRI) data to identify modular structures in brain functional networks. Modular structures are identified by using topological overlap matrix (TOM) elements in hierarchical clustering. We propose applying a new adaptive test built on the propor- tional odds model (POM) that can be applied to a high-dimensional setting, where the number of variables (p) can exceed the sample size (n) in addition to the usual $p < n$ setting. We applied our proposed methods to the ADNI data to test for associations between a genetic variant and either the whole brain functional network or its various subcomponents using various connectivity measures. We uncovered several modules based on the control cohort, and some of them were associated with the APOE4 variant.

Performance Assessment of High-dimensional Variable Estimation

Yanjia Yu, School of Statistics, University of Minnesota

Since model selection is ubiquitous in data analysis, reproducibility of statistical analysis demands a reality check of the employed model selection method no matter what label it may have in terms of good properties. Instability measures have been proposed for evaluating model selection uncertainty. However, low instability does not necessarily indicate that the selected model is trustworthy, since low instability can also arise when a certain method tends to select an overly parsimonious model. F and G -measure become increasingly popular for assessing variable selection performance in theoretical studies and simulation results. However, they are not available in practice. In this work, we propose an estimation ‘ methods in terms of model identification. We show that our approach provides reliable estimates of the true F and G measures of the selected models. This gives the data analyst a valuable tool to compare different model selection methods based on the data at hand. Extensive simulations are conducted to show its very good finite sample performance. We further demonstrate the application of our methods using several micro-array gene expression data sets.

Quantifying Publication Bias in Meta-Analysis

Lifeng Lin, Division of Biostatistics, University of Minnesota

Publication bias is a serious problem in systematic reviews and meta-analyses that can affect the validity and generalization of conclusions. Currently, the approaches to dealing with publication bias can be distinguished into two classes: the selection models and the funnel-plot-based methods. The selection models use weight functions to adjust the overall effect size estimate, and they are usually employed as sensitivity analyses to assess the potential impact of publication bias. The funnel-plot-based methods include the visual examination of funnel plot, the regression and rank tests, and the nonparametric trim and fill method. Although these approaches have been widely used in applications, measures for quantifying publication bias are seldom studied in the literature. Such a measure can be used as a characteristic of meta-analysis; also, it permits comparisons of publication biases across different meta-analyses. Egger's regression intercept may be considered as a candidate measure, but it lacks an intuitive interpretation. We introduce a new measure, skewness of standardized deviates, to quantify publication bias. This measure describes asymmetry of the collected study results. In addition, a new test for publication bias is derived based on the skewness. Large sample properties of the new measure are studied, and its performance is illustrated using simulations and three case studies.

An adaptive sum of powered correlation test for testing association between two matrices

Jason Xu, School of Statistics, University of Minnesota

Existing tests (e.g. RV test) may have low power in the presence of many nonassociated column pairs. For instance, Colantuoni et al. [2011] found that there was no global association between DNA variations and transcriptomic variations, although some indiviudal SNPs clearly was associated with expression levels of some individual genes. One possible reason for this surprising result is that a global SNP-expression association was hidden from many non-associated SNP-expression pairs. Motivated by this example, we propose a new method, called adaptive sum of powered correlation (aSPC) test, that can gain power by selecting large signals and reduce the effects of noise. We also point out the relationships between aSPC test with some other existing tests.

Gene- and pathway-based association tests for multiple traits with GWAS summary statistics

Il-Youp Kwak, Lillehei Heart Institute, University of Minnesota

To identify novel genetic variants associated with complex traits and to shed new insights on underlying biology, in addition to the most popular single SNP-single trait association analysis, it would be useful to explore multiple correlated (intermediate) traits at the gene- or pathway-level by mining existing single GWAS or meta-analyzed GWAS data. For this purpose, we present an adaptive gene-based test and a pathway-based test for association analysis of multiple traits with GWAS summary statistics. The proposed tests are adaptive at both the SNP- and trait-levels; that is, they account for possibly varying association patterns (e.g. signal sparsity levels) across SNPs and traits, thus maintaining high power across a wide range of situations. Furthermore, the proposed methods are general: they can be applied to mixed types of traits, and to Z-statistics or p-values as summary statistics obtained from either a single GWAS or a meta-analysis of multiple GWAS. Our numerical studies with simulated and real data demonstrated the promising performance of the proposed methods. The methods are implemented in R package aSPU, freely and publicly available at: `https://cran.r-project.org/web/packages/aSPU/`.