2018 Seminars

12/12/18

Fitting Competing Equations to the Data Using Smoothing Spline

Xinlian Zhang, PhDc

Abstract: A popular approach for flexible function estimation in nonparametric models is through spline smoothing using the general penalized likelihood method. In applying this method, one needs to pre- specify a set of equations which defines a penalty term and puts a soft constraint on the function to be estimated. A good choice of penalty, i.e., set of equations, is of key importance to the success of function estimation. In practice, defining the penalty functional is mostly based on expert knowledge of the system. However, for many dynamic systems there naturally exist more than one well-studied theory that explains the dynamics systems, i.e., there exist more than one set of sensible equations. To tackle this problem, we propose an approach that takes into consideration of all candidate penalties as well as the ambiguity in choosing among them. We take a fully Bayesian perspective, made use of the connection between penalized least squares and Bayesian estimation, and model the uncertainty of choosing penalty through introducing a mixture distribution as prior for parameters to be estimated. We also propose efficient sampling algorithm for making inference based on taking samples from posterior distribution.

Bio: Xinlian Zhang is currently a Ph.D. student in the Department of Statistics in the University of Georgia (UGA). She is interested in nonparametric function estimation and Bayesian spline smoothing analysis related problems in statistics. She is also interested in computational biology and application of statistical methods in biomedical analysis. Throughout her graduate studies, she has been collaborating with other labs working on small RNA related data and its role in transcriptional gene silencing and maintenance of genome stability in plants and nematodes.

12/05/2018

Probabilistic Projection of Carbon Emissions

Adrian Raftery, PhD

Abstract: The Intergovernmental Panel on Climate Change (IPCC) recently published climate change projections to 2100, giving likely ranges of global temperature increase for each of four possible scenarios for population, economic growth and carbon use. We develop a probabilistic forecast of carbon emissions to 2100, using a country-specific version of Kaya's identity, which expresses carbon emissions as a product of population, GDP per capita and carbon intensity (carbon per unit of GDP). We use the UN's probabilistic population projections for all countries, based on methods from our group, and develop a joint Bayesian hierarchical model for GDP per capita and carbon intensity in most countries. In contrast with opinion-based scenarios, our findings are statistically based using data for 1960–2010. We find that our likely range (90% interval) for cumulative carbon emissions to 2100 includes the IPCC's two middle scenarios but not the lowest or highest ones. We combine our results with the ensemble of climate models used by the IPCC to obtain a predictive distribution of global temperature increase to 2100. This is joint work with Dargan Frierson (UW Atmospheric Science), Richard Startz (UCSB Economics), Alec Zimmer (Upstart), and Peiran Liu (UW Statistics).

Bio: Dr. Adrian E. Raftery is a Professor of Statistics and Sociology at the University of Washington. He develops new statistical methods for problems in the social, environmental and health sciences. An elected member of the U.S. National Academy of Sciences, he was identified as the world's most cited researcher in mathematics for the decade 1995-2005 by Thomson-ISI. He has supervised 29 Ph.D. graduates, of whom 21 hold or have held tenure-track university faculty positions.

12/04/2018

Rank Based Inference for Clustered Data in Presence of Informative Intra-Cluster Group Size

Sandipan Dutta, PhD

Abstract: Rank based methods are useful non-parametric approaches of inference. Among them, the Wilcoxon rank-sum test is a popular nonparametric test for comparing two groups of independent observations. But, in many situations, observations may be correlated making the assumptions of the Wilcoxon rank-sum test invalid. One such scenario is a clustered data setting. In recent years, there have been renewed attempts in extending the Wilcoxon rank sum test for clustered data. However, in a clustered data setting, we are often faced with a situation where the group specific marginal distribution in a cluster depends on the number of observations in that group (i.e., the intra-cluster group size). In my presentation, we would talk about a novel extension of the rank-sum test for handling this complex situation. Moreover, we would also consider the scenario where the comparison of marginal group-specific distributions may not be enough in presence of some potentially useful covariables, such that ignoring the effects of these covariates can lead to incorrect inference for the group comparisons. To address this problem, we would discuss how to modify the rank-sum test into a covariate adjusted rank test by estimating the covariate effects through rank based estimation. We would demonstrate the utility of our proposed methods through a number of simulation studies and real-life datasets.

Bio: Dr. Sandipan Dutta is a postdoctoral research associate in the Department of Biostatistics and Bioinformatics at Duke University working in the research group of Professor Susan Halabi. He has obtained his Ph.D. in Biostatistics from the Department of Bioinformatics and Biostatistics at the University of Louisville under the supervision of Professor Somnath Datta. Prior to this, he obtained his masters in statistics from the Indian Institute of Technology Kanpur (IIT Kanpur), and his bachelors in statistics from the University of Calcutta, India. His research interests include developing methods for rank based inference of clustered data with complex features, time-to event analysis, censored data regression, prognostic modeling and identification of important biomarkers for clinical cancer data. His research works have been published in journals such as Biometrics, Statistics in Medicine, Journal of Statistical Computation and Simulation, Journal of Clinical Oncology among others.

11/28/18

Data versus Belief: Statistical Studies of Psychic Abilities

Jessica Utts, PhD

Abstract: After many years of investigating data in parapsychology (the study of possible psychic abilities), I have observed that belief and anecdotes often are given higher priority than data when people formulate conclusions about the possible existence of psychic phenomena. Unlike traditional frequentist methods of statistical inference, Bayesian methods allow the combination of data and prior beliefs to reach conclusions. Thus, the data from parapsychology provide a good example for comparing frequentist and Bayesian methods of making conclusions based on evidence. In this talk, I will present some of the data from experiments in parapsychology, and analyze it using both frequentist and Bayesian methods, illustrating how strong prior beliefs can be incorporated when we consider whether decision-makers will pay attention to data. This domain provides a good illustration of how Bayesian methods can be used in a real world setting, and how they allow people to disagree even when presented with a large amount of data. The talk concludes with an argument for why researchers in all areas should pay attention to the results of these experiments.

Bio: Dr. Jessica Utts is a Professor Emerita in the Department of Statistics at the University of California, Irvine, and was the 2016 President of the American Statistical Association. She received her PhD in Statistics from Penn State University and was a founding member of the UC Davis Statistics Department, where she served on the faculty there for many years before coming to UC Irvine. She has a long-standing interest in promoting

statistical literacy, and has published three statistics textbooks with that emphasis. She has been actively involved in writing and grading the AP Statistics exam since its inception in 1997, and has recently completed 5 years as the Chief Reader for the exam. In addition to statistics education her research involves applications of statistics to a variety of areas, most notably parapsychology, for which she has appeared on TV shows including Larry King Live, ABC Nightline and CNN Morning News.

11/27/18

Modern multivariate response regression with applications in genomics

Aaron Molstad, PhD

Abstract: In this two part talk, we will present new methods for fitting the multivariate response linear regression model motivated by applications in statistical genomics. In the first part, we introduce a new parameterization of the multivariate response linear regression model which we motivate through an "error-in-variables" data generating model. We propose a novel non-convex weighted residual sum of squares criterion which exploits this parameterization and admits a new class of penalized estimators. The optimization is solved with a proximal gradient descent algorithm. We use our method to study the association between copy number variations and gene expression in patients with glimoblastoma mulitforme, an aggressive brain cancer, collected by TCGA. In the second part of the talk, we propose a new multivariate response linear regression method for cross-tissue expression quantitative trait loci (eQTL) mapping in Genotype-Tissue Expression project (GTEx) data. Our method exploits that gene expression is dependent across tissue types and that eQTLs are often shared amongst multiple tissues. We propose a penalized maximum likelihood estimator and derive an efficient expectation-conditional-maximization algorithm for its computation. Our analysis of the GTEx data shows that our method can improve eQTL mapping substantially compared to methods which model expression separately tissue-by-tissue.

Bio: Dr. Aaron J. Molstad is a postdoctoral research fellow in the Biostatistics Program at the Fred Hutchinson Cancer Research Center in Seattle, WA. Previously, he received his Ph.D. from the School of Statistics at the University of Minnesota. Dr. Molstad's primary research interests are in multivariate analysis and numerical optimization, with an emphasis on developing model-based methods and open source software for statistical genetics and genomics. His past work has developed new methods for precision (inverse covariance) matrix estimation, classification with matrix and tensor-valued data, high-dimensional multivariate response linear regression, and survival analysis.

11/19/18

Interaction Feature Screening for Ultrahigh Dimensional Data

Guifang Fu, PhD

Abstract: Big data with ultrahigh dimensions has become increasingly important in perse scientific fields. For example, genome-wide association studies identify disease susceptibility loci by screening over half a million single-nucleotide polymorphisms (SNPs). Clinical study findings imply that complex diseases are very likely regulated by interactions among multiple genes (i.e., epistasis) rather than by one genetic variant within a single gene. However, selecting important interaction effects from an ultrahigh dimension of features is extremely challenging in terms of computational feasibility and statistical accuracy. Existing statistical approaches are either focused on marginal effects, have proven to be highly inaccurate for scenarios involving strong interactive but weak marginal effects, or are computationally infeasible for big data. In this presentation, I introduce a novel interaction screening procedure based on the joint cumulant correlation (JCM-SIS). The implementation of JCM-SIS does not require model specification or data type restriction for responses or predictors. We have performed four simulations under various conditions to comprehensively demonstrate that JCM-SIS is empirically accurate, robust, and computationally viable for features in ultrahigh dimensional

space. Numerical comparison indicates that JCM-SIS performs much better in a number of settings than two existing feature screening approaches. We successfully apply JCM-SIS to detect two-way interactions for 731,442 SNPs, a computational feat unprecedented in current literature. Additionally, we prove that JCM-SIS is theoretically sound and possesses strong sure screening consistency. In the discussion section, I will conclude by listing multiple future collaboration opportunities related to functional data analysis.

Bio: Dr. Guifang Fu is an Assistant Professor in the Department of Mathematics and Statistics at the Utah State University. She received her Ph.D. in Statistics from Pennsylvania State University. Her research has a dual focus in theoretical statistics methodologies and applied & computational statistics. She has led an independent, extramurally funded and nationally competitive research program while also actively collaborating on several multidisciplinary research teams. She specializes in developing state-of-the-art statistical methods to extract knowledge from data, advance the statistical theories that underlie these methods, and solve data-driven problems inspired by practical applications. She is highly interested in applying her methodologies to leading biomedical collaborations.

11/15/18

Propensity Score Weighting for Causal Inference with Multiple Treatments

Frank (Fan) Li

Abstract: Unconfounded comparisons of multiple groups are common in observational studies. Motivated from (1) an observational study comparing three medications (causal comparison) and (2) a racial disparity study (unconfounded descriptive comparison), we propose a unified framework, the balancing weights, for estimating causal effects with multiple treatments using propensity score weighting. These weights incorporate the generalized propensity score to balance the weighted covariate distribution of each treatment group, all weighted toward a common pre-specified target population. The class of balancing weights include several existing approaches such as inverse probability weights and trimming weights as special cases. Within this framework, we propose a class of target estimands based on linear contrasts and their corresponding nonparametric weighting estimators. We further propose the generalized overlap weights, constructed as the product of the inverse probability weights and the harmonic mean of the generalized propensity scores. The generalized overlap weights correspond to the target population with the most overlap in covariates between treatments, similar to the population in equipoise in randomized clinical trials. These weights are bounded and thus bypass the problem of extreme propensities. We show that the generalized overlap weights minimize the total asymptotic variance of the nonparametric estimators for the pairwise contrasts within the class of balancing weights. We consider two balance check criteria and propose a new sandwich variance estimator for estimating the causal effects with generalized overlap weights. We apply these methods to (1) study the causal effect of three anti-coagulants on patient's mortality and (2) to estimate the racial disparities in medical expenditure. The operating characteristics of the new weighing method is further illustrated by simulations.

Bio: Frank (Fan) Li is a Ph.D. candidate in the Department of Biostatistics and Bioinformatics at Duke University, and a student affiliate at the Duke Clinical Research Institute. His primary research interests include causal inference methods applied to observational studies and pragmatic trials. In his doctoral dissertation, he studied propensity score methods for difference-in-differences and multiple treatments. He is also an active member of the Biostatistics and Study Design core in the NIH Collaboratory of Pragmatic Clinical Trials, established to oversee the statistical issues of ongoing demonstration projects.

11/07/2018

Surveys and Big Data for Estimating Brand Lift

Tim Hesterberg, PhD

Abstract: Google Brand Lift Surveys estimates the effect of display advertising using surveys. Challenges include imperfect A/B experiments, response and solicitation bias, discrepancy between intended and actual treatment, comparing treatment group users who took an action with control users who might have acted, and estimation for different slices of the population. We approach these issues using a combination of inpidual-study analysis and meta-analysis across thousands of studies. This work involves a combination of small and large data - survey responses and logs data, respectively. There are a number of interesting and even surprising methodological twists. We use regression to handle imperfect A/B experiments and response and solicitation biases; we find regression to be more stable than propensity methods. We use a particular form of regularization that combines advantages of L1 regularization (better predictions) and L2 (smoothness). We use a variety of slicing methods, that estimate either incremental or non-incremental effects of covariates like age and gender that may be correlated. We bootstrap to obtain standard errors. In contrast to many regression settings, where one may either resample observations or fix X and resample Y, here only resampling observations is appropriate.

Bio: Dr. Tim Hesterberg is a Senior Statistician at Google. That means old. Before that he attempted jobs at an electric utility, in academia, and in software. He received his Ph.D. in Statistics from Stanford University, where he played a lot of volleyball. He wrote Mathematical Statistics with Resampling and R; he helped write the ASA Guidelines for Undergraduate Statistics Programs, so he could tell teachers how to teach. Now he'll tell you how to analyze data!

7/03/18

Bayesian Approaches to Dynamic Model Selection

Michele Guindani, PhD

Abstract: In many applications, investigators monitor processes that vary in space and time, with the goal of identifying temporally persistent and spatially localized departures from a baseline or “normal" behavior. In this talk, I will first discuss a principled Bayesian approach for estimating time varying functional connectivity networks from brain fMRI data. Dynamic functional connectivity, i.e., the study of how interactions among brain regions change dynamically over the course of an fMRI experiment, has recently received wide interest in the neuroimaging literature. Our method utilizes a hidden Markov model for classification of latent neurological states, achieving estimation of the connectivity networks in an integrated framework that borrows strength over the entire time course of the experiment. Furthermore, we assume that the graph structures, which define the connectivity states at each time point, are related within a super-graph, to encourage the selection of the same edges among related graphs. Then, I will propose a Bayesian nonparametric model selection approach with an application to the monitoring of pneumonia and influenza (P&I) mortality, to detect influenza outbreaks in the continental United States. More specifically, we introduce a zero-inflated conditionally identically distributed species sampling prior which allows borrowing information across time and to assign data to clusters associated to either a null or an alternate process. Spatial dependences are accounted for by means of a Markov random field prior, which allows to inform the selection based on inferences conducted at nearby locations. We show how the proposed modeling framework performs in an application to the P&I mortality data and in a simulation study, and compare with common threshold methods for detecting outbreaks over time, with more recent Markov switching.

Bio: Dr. Guindani is a Professor in the Department of Statistics, University of California, Irvine. Before joining UCI, he has held faculty positions in the Department of Biostatistics, University of Texas MD Anderson Cancer Center and the Department of Mathematics and Statistics at the University of New Mexico. He has received his Ph.D. in Statistics from Università Bocconi, Milan, Italy in Spring 2005. He is currently a Co-Editor for Bayesian Analysis, the official journal of the International Society for Bayesian Analysis (ISBA) and he has been nominated as Editor-in-Chief of the same journal from January 2019 to December 2021. He is also an Associate Editor for Biometrics.

06/06/2018

Digging into the Biology Complex Traits

Rany Salem, PhD

Abstract: The emergence of Genome-wide association studies (GWAS) ushered a paradigm shift in human genetics and genetic epidemiology research in terms of the study designs and methodologies researchers could exploit to detect genetic factors contributing to a complex traits and disease. GWAS have allowed investigators to probe the genetic architecture and identify loci associated with thousands of human traits and diseases. In this talk, I will provide an overview of the lab's research focus and describe results of a GWAS study of Diabetic Kidney Disease to illustrate the utility, challenges and limitations of such studies. Next, potentials solutions to these limitations are presented, including leveraging inpidual GWAS data available in central biorepositories (e.g. dbGaP) and GWAS summary statistics. I will also discuss recent work to predict weight gain using metabolomics data. Finally, I will briefly describe current limitations, opportunities and new directions in human genetics and genetic epidemiology.

Bio: Dr. Rany Salem is an Assistant Professor in Department of Family Medicine and Public Health. His research interests focus on application of statistical and epidemiological methods to understand the genetic architecture of complex traits and disease, including diabetes and diabetic complications, metabolic syndrome, cardiovascular and renal disease, and anthropometric and growth traits. He is interested in methodological questions in human genetics and leverage publically available datasets to explore genetics questions at scale. He received his Ph.D. in Public Health from the UC San Diego/SDSU Joint Doctoral Program in 2009 under the mentorship of Drs. Nik Schork and Dan O'Connor. During his doctoral studies, he completed the UC San Diego Genetics Training Program in 2007. After receiving his Ph.D., he worked first as a postdoctoral fellow and then as a Senior Research Fellow at the Broad Institute, Boston Children's Hospital, Harvard Medical School, where he focused on genetic epidemiology studies with an emphasis on statistical genetic methodology and analysis of large epidemiologic datasets. During his postdoc training, he was awarded an NIH K99 Pathway to Independence Award from NHLBI.

06/01/2018

Controlling Epidemics II: Challenges and Opportunities for Quantitative Scientists

Victor DeGruttola PhD, ScD

Abstract: Recent developments in biomedical science, such as those in molecular epidemiology and surveillance, vaccinology, and antimicrobial treatment have shown promise in improving design and evaluation of interventions to control epidemic and endemic diseases. These methods help to address challenges resulting from the complex dependencies that arise in data from clinical studies; the dependencies reflect the spread of communicable diseases along contact networks. Both randomized and observational studies often collect data on HIV incidence in different subpopulations, risk behavior, and viral genetic sequences. This information can aid not only in design of studies but also in improving efficiency of analysis and scale-up of results. These can be complicated by interference that arises when the treatment of one unit can affect the outcome of another--a situation likely to arise with outcomes that may depend on social interactions, such those that transmit infectious diseases. To address the issue of scale-up of results from clinical trials subject to interference across randomized units, we focus on an estimand defined as the difference in the outcome that one would observe if the treatment were provided to all clusters compared to that outcome if treatment were provided to none – referred as the overall treatment effect. In trials of infectious disease prevention, the randomized treatment effect estimate will be attenuated relative to this overall treatment effect if a fraction of the exposures in the treatment clusters come from individuals who are outside these clusters. We leverage epidemic models to infer the way in which a given level of interference affects the incidence of infection in

clusters. This leads naturally to an estimator of the overall treatment effect that is easily implemented using existing software. In another setting where interference is likely and—Ebola outbreaks--we propose and demonstrate properties of a novel design; this design is most relevant when a proof-of-principle vaccine trial has been conducted, but questions remain about the effectiveness of different possible modes of implementation. Our goal for these studies is not only to generate information about intervention effects but also to provide public health benefit. To do so, we leverage information about contact networks – in particular the degree of connection across randomized units obtained at study baseline – and develop a novel class of connectivity-informed cluster trial designs. We investigate the performance of these designs in terms of epidemic control outcomes (time to end of epidemic and cumulative incidence) and power to detect intervention effect, by simulating vaccination trials during an SEIR-type epidemic outbreak using a network-structured agent-based model.

Bio: Dr. Victor De Gruttola has spent the past 30 years working with junior colleagues and in collaboration with clinical and laboratory investigators to develop and apply methods for advancing the HIV prevention and treatment research agendas. He also has managed large projects devoted to improving the public health response to the AIDS epidemic, both within the US and internationally. The aspects of the HIV epidemic on which he has worked include transmission and natural history of infection with the Human Immunodeficiency Virus (HIV), as well as investigation of antiretroviral treatments, including the development and consequences of resistance to them. The broad goals of his research have included developing treatment strategies that provide durable virologic suppression while preserving treatment options after failure, and evaluating the community-level impact of packages of prevention interventions, including antiviral treatment itself. He served as the Director of the Statistics and Data Analysis Center of the Adult Project of the AIDS Clinical Trials Group during the period in which highly active antiretroviral treatment was developed, and was instrumental in designing and analyzing studies of the best means of providing such therapy. He has also served as the Co-PI (with Max Essex) for a cluster-randomized trial of an HIV combination prevention program in Botswana. His methods research activity is focused on HIV prevention research, especially with regard to the development of methods for analyses of sexual contact networks, for viral genetic linkage analyses in the presence of missing data, and for improving validity and efficiency of analyses of HIV prevention trials.

05/02/2018

Meta-Analysis of Odds Ratios With Missing Counts Estimated using Kaplan-Meier Curves

Shemra Rizzo, PhD

Abstract: A typical random effects meta-analysis of odds-ratios assumes binomially distributed numbers of events in a treatment and control group and requires the number of events (i.e. deaths) and non-events (i.e. survivors) to be extracted from published papers. These data are often not available in the publications due to loss to follow-up. When the Kaplan-Meier (KM) survival plot is available, it is common practice to extract the survival probability from the plot and multiply it by the baseline sample size to infer the number of deaths and survivors. The naive approach to meta-analysis introduces these estimates as real extracted data; the results are hence over-certain and potentially inaccurate. Furthermore, accounting for the uncertainty introduced from these calculations is difficult as KM curves are typically published without variance information. We propose a model to incorporate the uncertainty associated with the estimation of the missing counts that uses summary statistics for the follow-up times. Furthermore, accounting for the uncertainty of the estimation is equivalent to a reduction of each study's sample size. A simulation study shows that our model outperforms the naive approach in terms of the coverage of the 95% confidence interval. We use real and simulated data to illustrate our model.

Bio: Dr. Shemra Rizzo is assistant professor at UC Riverside and currently serving as Vice-President of Academic Affairs for the Southern California Chapter of the American Statistical Association. She obtained her masters degree in statistics and operations research from the University of North Carolina - Chapel Hill and her PhD in biostatistics from UCLA.

04/11/2018

Estimating Network Properties: Applications to Sexual History Data

Ravi Goyal, PhD

Abstract: Analysis of sexual history data intended to describe sexual networks presents many challenges arising from the fact that most surveys collect information on only a very small fraction of the population of interest. In addition partners are rarely identified and responses are subject to reporting biases. Typically each network property of interest, such as mean number of sexual partners for males or females, is estimated independently of other network properties. There is, however, a complex relationship among networks properties; and knowledge of these relationships can aid in addressing concerns mentioned above. This talk will present a method that leverages the relationships among network properties when making inferences about network features of interest. The method ensures that inference on network properties is compatible with an actual network. The talk will present simulation results which demonstrate that use of this method can improve estimates in settings where there is uncertainty that arises both from sampling and from systematic reporting bias compared to currently available approaches. The talk with conclude with applying the method to estimate network properties using data from the Chicago Health and Social Life Survey.

Bio: Dr. Ravi Goyal (Ph.D., Biostatistics, Harvard School of Public Health) is a statistician at Mathematica Policy Research where he focuses on developing and applying statistical network analysis methodology to improve and evaluate public programs. As a graduate student and research associate at Harvard University, his research focused on developing network sampling methods that capture uncertainties in network structure and applying these methods to analyze and model HIV epidemic data. During his employment as an applied mathematician at National Security Agency, he gained field experience (deployed to Iraq) and experience with real world complex datasets that included geospatial, longitudinal, and social network data.

04/02/2018

Bayesian regression for group testing data

Joshua M. Tebbs

Abstract: Group testing involves pooling inpidual specimens (e.g., blood, urine, swabs, etc.) and testing the pools for the presence of a disease. When inpidual covariate information is available (e.g., age, gender, number of sexual partners, etc.), a common goal is to relate an inpidual's true disease status to the covariates in a regression model. Estimating this relationship is a nonstandard problem in group testing because true inpidual statuses are not observed and all testing responses (on pools and on inpiduals) are subject to misclassification arising from assay error. Previous regression methods for group testing data can be inefficient because they are restricted to using only initial pool responses and/or they make potentially unrealistic assumptions regarding the assay accuracy probabilities. To overcome these limitations, we propose a general Bayesian regression framework for modeling group testing data. The novelty of our approach is that it can be easily implemented with data from any group testing protocol. Furthermore, our approach will simultaneously estimate assay accuracy probabilities (along with the covariate effects) and can even be applied in screening situations where multiple assays are used. We apply our methods to group testing data collected in Iowa as part of statewide screening efforts for chlamydia.

Bio: Dr. Joshua Tebbs is Professor in the Department of Statistics in the College of Arts and Sciences at University of South Carolina. He received his BS in Mathematics (1995) and MS in Statistics (1997) from University of Iowa and his PhD in Statistics (2000) from North Carolina State University. He is a Fellow of the American Statistical Association. His research involves the development of statistical methods for categorical data, primarily binary response data that are observed in pools (group testing), and for constrained inference problems motivated by biomedical and public health applications. His research has been funded by two R01

grants from the National Institutes of Health (NIH), he routinely serves on NIH and NSF review panels, and he has advised eight PhD students. His current work is aimed at surveillance and identification for multiple diseases in group testing applications, motivated by nationwide screening activities for chlamydia and gonorrhea.

02/27/2018

Optimal treatment allocations in space and time for online control of an emerging infectious disease

Eric Laber, PhD

Abstract: A key component in controlling the spread of an epidemic is deciding where, when, and to whom to apply an intervention. We develop a framework for using data to inform these decisions in real-time. We formalize a treatment allocation strategy as a sequence of functions, one per treatment period, that map up-to-date information on the spread of an infectious disease to a subset of locations where treatment should be allocated. An optimal allocation strategy optimizes some cumulative outcome, e.g., the number of uninfected locations, the geographic footprint of the disease, or the cost of the epidemic. Estimation of an optimal allocation strategy for an emerging infectious disease is challenging because spatial proximity induces interference among locations, the number of possible allocations is exponential in the number of locations, and because disease dynamics and intervention effectiveness are unknown at outbreak. We derive a Bayesian online estimator of the optimal allocation strategy that combines simulation-optimization with Thompson sampling. The proposed estimator performs favorably in simulation experiments. This work is motivated by and illustrated using data on the spread of white-nose syndrome, a highly fatal infectious disease devastating bat populations in North America.

Bio: Dr. Eric Laber is an Associate Professor in Department of Statistics in North Carolina State University. His major research areas are causal inference, non-regular asymptotics, optimization, and reinforcement learning. The primary application areas include precision medicine, artificial intelligence, adaptive conservation, and the management of infectious diseases.

01/19/2018

Statistical methods for high-throughput genomic data

Zhixiang Lin, PhD

Abstract: In the first part of the talk, a dimension reduction method will be introduced where we extend Principal Component Analysis to propose AC-PCA for simultaneous dimension reduction and Adjustment for Confounding variation. We show that AC-PCA can adjust for variations across inpidual donors present in a human brain dataset. For gene selection purposes, we extend AC-PCA with sparsity constraints, and propose and implement an efficient algorithm. The second part of the talk will be focused on clustering methods in single cell genomics. In single cell genomics, it is technically challenging to obtain chromatin accessibility and gene expression data for the same cell. We have developed a computational approach to this problem, where a model-based clustering method is proposed to match cell sub-populations in these two data types. We also demonstrate that using one data type can guide clustering of the other data type. Our proposed Bayesian model accounts for the stochasticity due to biological and technical effects. Last, methodologies motivated by spatial temporal modeling of gene expression dynamics during human brain development will be briefly discussed.

Bio: Dr. Zhixiang Lin studied biological sciences at Tsinghua University (BS, 2010), computational biology & bioinformatics and statistics at Yale University (PhD, 2015). He is a postdoctoral scholar at Stanford University, Department of Statistics since 2015. His major research area is statistical genetics/genomics and

computational biology. His work has been published in prestigious journals such as PNAS, Biometrics, Annals of Applied Statistics and Cell.

01/12/2018

A semi-supervised approach for predicting cell type/tissue specific functional consequences of non-coding variation using massively parallel reporter assays

Zihuai He, PhD

Abstract: Predicting the functional consequences of genetic variants is a challenging problem, especially for variants residing in non-coding regions. Projects such as ENCODE and Roadmap Epigenomics make available various epigenetic features, including histone modifications and chromatin accessibility, genome-wide in over a hundred different tissues and cell types. Meanwhile, recent developments in high-throughput assays to assess the functional impact of variants in regulatory regions (e.g. massively parallel reporter assays - MPRA, CRISPR/Cas9-mediated in situ saturating mutagenesis) can lead to the generation of high quality data on the functional effects of selected variants. We propose a semi-supervised approach, referred to as GenoNet, to jointly utilize experimentally confirmed regulatory variants (labeled variants), millions of unlabeled variants genome-wide, and more than a thousand cell type/tissue specific functional annotations on each variant to predict functional consequences of non-coding genetic variants. Through the application to several experimental datasets, we demonstrate that the proposed method significantly improves prediction accuracy compared to existing functional prediction methods, both at the organism level and at the tissue/cell type level. We further show that eQTLs and dsQTLs in specific tissues tend to be substantially more enriched among variants with high GenoNet scores, and how the GenoNet scores can be used to map regulatory variants in regions of interest, evaluate 3C interaction variants and aid in the discovery of disease associated genes through an integrative analysis of lipid phenotypes using a Metabochip dataset on 12,281 inpiduals.

Bio: Dr. Zihuai He received his Ph.D. in Biostatistics at the University of Michigan, and BS (Bachelor of Science) at Tsinghua University in China. He is currently a post-doctoral research scientist in the Department of Biostatistics at Columbia University. His research has been concentrated in the area of statistical genetics and integrative analysis of omics data. There have been 11 peer-reviewed journal publications generated from his work published in prestigious journals of genetics and statistics, such as The American Journal of Human Genetics, Journal of the American Statistical Association, and Biometrics. He has developed three R packages with efficient computational techniques that facilitate integrative analysis in a broad range of genomic study designs such as longitudinal studies, family studies and meta-analysis of multiple sequencing studies. At Columbia, he also collaborates with researchers in the GTEx Consortium for gene expression studies.

01/05/2018

Joint modeling of longitudinal functional feature and discrete time-to-event

Ling Ma, PhD

Abstract: In longitudinal studies, it is often of interest to investigate how the functional feature of a marker's measurement process is associated with the event time of interest. We make use of B-splines to smoothly approximate the infinite dimensional functional data and propose a joint model of the longitudinal functional feature and the time to event. The proposed approach also allows for prediction of survival probabilities for future subjects based on their available longitudinal measurements and a fitted joint model. We illustrate our proposals on a prospective pregnancy study, namely Oxford Conception Study, where hormonal measurements of luteinizing hormone which is an important biomarker of ovulation is available. A joint modeling approach using functional analytic approach and discrete survival modeling was used to assess whether the functional feature of hormonal measurements, such as the curvature of the hormonal profile is associated with time to pregnancy.

Bio: Dr. Ling Ma received her PhD in Statistics from University of Missouri, Columbia in 2014. She then worked at the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) as a postdoctoral fellow for two years before joining Clemson University as an assistant professor. Dr. Ma's primary methodological research interests are survival analysis with special emphasis on interval-censored data and panel count data, joint modeling of longitudinal and time-to-event data. Dr. Ma has worked on statistical methods with applications to reproductive and environmental epidemiology, cancer, HIV, etc.