Statistical and Computational Theory and Methodology for Big Data Analysis
Feb 9-Feb 14, 2014

MEALS

Breakfast (Buffet): 7:00-9:00am, Sally Borden Building, Monday-Friday
Lunch (Buffet): 11:30am-1:30pm, Sally Borden Building, Monday-Friday
Dinner (Buffet): 5:30pm-7:30pm, Sally Borden Building, Sunday-Thursday

SCHEDULE

Sunday

16:00 Check-in begins
17:30-19:30 Buffet Dinner
20:00 Informal gathering

Monday, February 10

7:00-8:45 Breakfast
8:45-9:00 Introduction and Welcome by BIRS Station Manager

Morning Session I, Chair: Ming-Hui Chen
9:00-9:05 Workshop Opening: Ming-Hui Chen
9:05-9:25 Special issues: Peihua Qiu, Heping Zhang
9:25-9:55 ASA video

9:55-10:25 Coffee Break

Morning Session II, Chair: Radu Craiu
10:25-11:05 Hongzhe Li: *Microbiome, Metagenomics and High-dimensional Compositional Data Analysis*
11:05-11:45 Heping Zhang: *Tree-based Rare Variants Analyses*
11:45-11:55 Floor Discussion

12:00-13:00 Lunch

13:00-13:50 Guided Tour of the Banff Center

13:50 Group photo

Afternoon Session I, Chair: Faming Liang

14:00-14:40 Marc A. Suchard: *When multi-core statistical computing fails for massive sample sizes …*

14:40-15:20 Minge Xie: *A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data*

15:20-15:30 Floor Discussion

15:30-15:50 Coffee Break

Afternoon Session II, chair: Chuanhai Liu

15:50-16:30 Ping Li: *BigData: Efficient Search and Learning using Sparse Random Projections and Probabilistic Hashing*

16:30-17:10 Christophe Andrieu: *Uniform Ergodicity of the Iterated Conditional SMC and Geometric Ergodicity of Particle Gibbs Samplers*

17:10-17:50 Lingsong Zhang: *Scale-Space Inference with Application on Spatial Clustering Detection*

17:50-18:00 Floor Discussion

18:00-19:30 Dinner

Tuesday, February 11

7:00-8:30 Breakfast

Morning Session I, Chair: Jianhua Huang

8:30-9:10 Peihua Qiu: *On Nonparametric Profile Monitoring*

9:10-9:50 Hongtu Zhu: *Functional Data Analysis of Imaging Data*

9:50-10:00 Floor Discussion

10:00-10:30 Coffee Break

Morning SessionG II, Chair: Marc Suchard

10:30-11:10 Jian Zhang: *High-dimensional Inference in Magnetoencephalographic Neuroimaging*

11:10-11:50 Momiao Xiong: *Classification Analysis of Big Image Data*

11:50-12:00 Floor Discussion

12:00-13:30 Lunch

Afternoon Session I, Chair: Bo Li

13:30-15:10 Faming Liang & Chuanhai Liu: *Recent Developments of Iterative Monte Carlo Methods for Big Data Analysis*
15:10-15:20 Floor Discussion

15:20-15:50 Coffee Break

Afternoon Session II, Chair: Guanghua Xiao

15:50-17:30 Jun Yan: *Recent Software Development for Big Data Analysis*
17:30-17:40 Floor Discussion

17:40-19:30 Dinner

Wednesday, February 12

7:00-8:30 Breakfast

Morning Session I, Chair: Andrieu Christophe
8:30-9:10 Valen Johnson: *Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings*

9:10-9:50 Linglong Kong: *Quantile regression in Variable Screening*
9:50-10:00 Floor Discussion

10:00-10:30 Coffee Break

Morning Session II, Chair: Lynn Kuo
10:30-11:10 Yingnian Wu: *What Is Beyond Sparse Coding?*
11:10-11:50 Xiao Wang: *Functional Regression and Image Regression*
11:50-12:00 Floor Discussion

12:00-13:30 Lunch

Afternoon Session I, Chair: Yuguo Chen

13:30-14:10 Xiaotong Shen: *Sentiment Analysis*
14:10-14:50 Xiaojing Wang: *A Bayesian Approach to Subgroup Identification*
14:50-15:30 Guanghua Xiao: *Detection of tumor driver genes using a fully integrated Bayesian approach*
15:30-15:40 Floor Discussion

15:40-16:00 Coffee Break

Afternoon Session II, Chair: Ruslan Salakhutdinov

16:00-16:40 Elizabeth Schifano: *Online Updating of Statistical Inference in the Big Data Setting*
16:40-17:20 Nan Lin: *Statistical Aggregation in Massive Data Environment*
17:20-17:30 Floor Discussion

17:30-19:30 Dinner

Thursday, February 13

7:00-8:30 Breakfast

Morning Session I, Chair: Farshad Farshidfar

8:30-9:10 Hongyu Zhao: *Detecting Genetic Association Signals Leveraging Network Information*
9:10-9:50 Zhang Zhang: *Biocuration in the Era of Big Data*
9:50-10:00 Floor Discussion

10:00-10:30 Coffee Break

Morning Session II, Chair: Kun Chen

10:30-11:10 Xin Gao: *Poly(A) motif Prediction Using Spectral Latent Features from Human DNA Sequences*

11:10-11:50 Matthias Katzfuss: *Statistical Inference for Massive Distributed*

Spatial Data Using Low-Rank Models

11:50-12:00 Floor Discussion

12:00-13:30 Lunch

Afternoon Session I, Chair: Alexander Shestopaloff

13:30-14:10 Ruslan Salakhutdinov: *Annealing Between Distributions by Averaging Moments*
14:10-14:50 Alex Shestopaloff: *MCMC for non-Linear State Space Models Using Ensembles of Latent Sequences*
14:50-15:30 Xiaoyi Min: *The Screening and Ranking Algorithm for Change-Points Detection in Multiple Samples*
15:30-15:40 Floor Discussion

15:40-16:00 Coffee Break

Afternoon Session II, Chair: Paul Kvam

16:00-16:40 Lee Dicker: *Variance Estimation for High-Dimensional Linear Models*
16:40-17:20 Philip Gautier: *D&R for Large Complex Data: Likelihood Modeling for Logistic Regression*
17:20-17:30 Floor Discussion

17:30-19:30 Dinner

Friday, February 14

7:00-8:30 Breakfast

8:30-12:00 Free discussion

12:00-13:30 Lunch

Checkout by 12 noon

ABSTRACT

Monday, February 10

Microbiome, Metagenomics and High-dimensional Compositional Data Analysis

Hongzhe Li

Department of Biostatistics, University of Pennsylvania., USA

Next-generation sequencing technologies allow 16S ribosomal RNA gene surveys or whole metagenome shotgun sequencing in order to characterize taxonomic and functional compositions of gut microbiomes. The outputs from such studies are short sequence reads derived from a mixture of genomes of different species in a given microbial community. We first present a brief overview of the statistical methods we used for 16S rRNA data analysis. We then introduce a multi-sample model-based method to quantify the bacterial compositions based on shotgun metagenomics using species-specific marker genes. The resulting data are high-dimensional compositional data, which complicate many of the downstream analyses. We introduce the GLMs with linear constraint on regression parameters in order to identify the bacterial taxa that are associated clinical outcomes and a composition-adjusted thresholding procedure to estimate correlation network from compositional data. We demonstrate the methods using two on-going gut microbiome studies at the University of Pennsylvania.

Tree-based Rare Variants Analyses

Chi Song and Heping Zhang*

Department of Biostatistics, Yale School of Public Health, Yale University, USA.

Since the development of next generation sequencing (NGS) technology, researchers have been extending their efforts on genome-wide association studies (GWAS) from common variants to rare variants to find the missing inheritance. Although various statistical methods have been proposed to analyze rare variants data, they generally face difficulties for complex disease models involving multiple genes. In this paper, we propose a tree-based method that adopts a non-parametric disease model and is capable of exploring gene-gene interactions. We found that our method outperforms the sequence kernel association test (SKAT) in most

of our simulation scenarios, and by notable margins in some cases. By applying the tree-based method to the Study of Addiction: Genetics and Environment (SAGE) data, we successfully detected gene CTNNA2 and its 44 specific variants that increase the risk of alcoholism in women. This gene has not been detected in the SAGE data. Post hoc literature search also supports the role of CTNNA2 as a likely risk gene for alcohol addiction. This finding suggests that our tree-based method can be effective in dissecting genetic variants for complex diseases using rare variants data.

When Multi-Core Statistical Computing Fails for Massive Sample Sizes …

Marc A. Suchard

Department of Biomathematics, Biostatistics and Human Genetics

University of California Los Angeles

Much of statistical computing is memory-bandwidth limited, not floating-pointing operation throughput limited as commonly assumed. This often restricts the utility of multi-core computing techniques to improve statistical estimation run-time. I explore this conundrum in inference tools for a massive Bayesian model of sea-surface temperatures across the global. I describe approaches for computing the data likelihood that exploit fine-scale parallelization for potential scalability to real-time satellite surveillance data. These simple algorithmic changes open the door on using advancing computing technology involving many-core architectures. These architectures provide significantly higher memory-bandwidth and inexpensively afford order-of-magnitude run-time speed-ups.

A Split-and-Conquer Approach for Analysis of Extraordinarily Large Data

Minge Xie

Department of Statistics, Rutgers University

If there are extraordinarily large data, too large to fit into a single computer or too expensive to perform a computationally intensive data analysis, what should we do? To deal with this problem, we propose in this paper a “split-and-conquer'' approach and illustrate it using several computationally intensive penalized regression methods, along with a theoretical support. Consider a regression setting of generalized linear models with n observations and p covariates, in which n is extraordinarily large and p is either bounded or goes to ∞ at a certain rate of n. We propose to randomly split the data of size n into K subsets of size $O(n/K)$. For each subset of data, we perform a penalized regression analysis and the results from each of

the K subsets are then combined to obtain an overall result. We show that under mild conditions the combined overall result still retains desired properties of many commonly used penalized estimators, such as the model selection consistency and asymptotic normality. When K is well controlled, we also show that the combined result is asymptotically equivalent to the result of analyzing the entire data all at once (assuming that there is a super computer that could carry out such an analysis). In addition, when a computational intensive algorithm is used in the sense that its computing expense is at the order of $O(n^a p^b)$, $a > 1$ and $b \geq 0$, we show that the split-and-conquer approach can substantially reduce computing time and computer memory requirement. Furthermore, we demonstrate that the approach has an inherent advantage of being more resistant to false model selections caused by spurious correlations. Similar to what reported in the literature, we can establish an upper bound for the expected number of falsely selected variables and a lower bound for the expected number for truly selected variables. The proposed methodology is illustrated numerically using both simulation and real data examples.

BigData: Efficient Search and Learning Using Sparse Random Projections and Probabilistic Hashing

Ping Li

Department of Statistics, Rutgers University

Modern applications of search and learning have to deal with datasets with billions of examples in billion or even billion square dimensions (e.g., text documents represented by high-order n-grams). In this talk, we will first present the use of very sparse random projections (Li, Hastie, Church, KDD 2006) for learning with high-dimensional data. It is evident that the projection matrix can be extremely sparse (e.g., 0.1% or less nonzeros) without hurting the learning performance. For binary sparse data (which are common in practice), however, b-bit minwise hashing (Li and Konig, Communications of the ACM 2011) turns out to be much more efficient than random projections. In addition, the recent development of one-permutation hashing (Li, Owen, Zhang, NIPS 2012) substantially reduced the processing time of (b-bit) minwise hashing, from (e.g.,) 500 permutations to merely one. There are many other exciting new progresses in the basic research of random projections and hashing, for example, the new work on sign Cauchy random projections for approximating chi-square distances (Li, Samorodnitsky, Hopcroft, NIPS 2013) and the work on using stable random projections for very fast and accurate compressed sensing (Li, Zhang, Zhang, 2013).

Uniform Ergodicity of the Iterated Conditional SMC and Geometric Ergodicity of Particle Gibbs samplers

Christophe Andrieu

School of Mathematics, University of Bristol, Bristol BS8 1TW, UK

We establish quantitative bounds for rates of convergence and asymptotic variances for iterated conditional sequential Monte Carlo (i-cSMC) Markov chains and associated particle Gibbs samplers. Our main findings are that the essential boundedness of potential functions associated with the i-cSMC algorithm provide necessary and sufficient conditions for the uniform ergodicity of the i-cSMC Markov chain, as well as quantitative bounds on its (uniformly geometric) rate of convergence. This complements more straightforward results for the particle independent Metropolis--Hastings (PIMH) algorithm. Our results for i-cSMC imply that the rate of convergence can be improved arbitrarily by increasing N, the number of particles in the algorithm, and that in the presence of mixing assumptions, the rate of convergence can be kept constant by increasing N linearly with the time horizon. Neither of these phenomena are observed for the PIMH algorithm. We translate the sufficiency of the boundedness condition for i-cSMC into sufficient conditions for the particle Gibbs Markov chain to be geometrically ergodic and quantitative bounds on its geometric rate of convergence. These results complement recently discovered, and related, conditions for the particle marginal Metropolis--Hastings (PMMH) Markov chain. This is joint work with Anthony Lee and Matti Vihola.

Scale-Space Inference with Application on Spatial Clustering Detection

Lingsong Zhang

Department of Statistics, Purdue University

A novel multi-resolution cluster detection (MCD) method is proposed to identify irregularly shaped clusters in space. Multi-scale test statistic on a single cell is derived based on likelihood ratio statistic for Bernoulli sequence, Poisson sequence and Normal sequence. A neighborhood variability measure is defined to select the optimal test threshold. The MCD method is compared with single scale testing methods controlling for false discovery rate and the spatial scan statistics using simulation and f-MRI data. The MCD method is shown to be more effective for discovering irregularly shaped clusters, and the implementation of this method does not require heavy computation, making it suitable for cluster detection for large spatial data.

Tuesday, February 11

On Nonparametric Profile Monitoring

Peihua Qiu

Department of Biostatistics, University of Florida

Quality of a process is often characterized by the functional relationship between a response and one or more predictors. Profile monitoring is for checking the stability of this relationship over time. In the literature, most existing control charts are for monitoring parametric profiles, and they assume that within-profile observations are independent of each other, which is often invalid. This talk presents some of our recent research on nonparametric profile monitoring when within-profile data are correlated. We will also briefly describe the problems of online image monitoring and dynamic disease screening that are closely related to profile monitoring.

Functional Data Analysis of Imaging Data

Hongtu Zhu

Department of Biostatistics, University of North Carolina at Chapel Hill

Motivated by recent work on studying massive imaging data in various neuroimaging studies,our group proposes several classes of spatial regression models including spatially varying coefficient models, spatial predictive Gaussian process models, tensor regression models, and Cox functional linear regression models for the joint analysis of large neuroimaging data and clinical and behavioral data. Our statistical models explicitly account for several stylized features of neuorimaging data: the presence of multiple piecewise smooth regions with unknown edges and jumps and substantial spatial correlations. We develop some fast estimation procedures to simultaneously estimate the varying coefficient functions and the spatial correlations. We systematically investigate the asymptotic properties (e.g., consistency and asymptotic normality) of the multiscale adaptive parameter estimates. Our Monte Carlo simulation and real data analysis have confirmed the excellent performance of our models in different applications.

High-dimensional Inference in Magnetoencephalographic Neuroimaging

Jian Zhang

University of Kent, UK

Estimation of a high-dimensional time-varying coefficient model on the basis of spatially correlated observations is one of challenging problems in statistics. Our study was motivated by source localization problem in magnetoencephalographic (MEG) neuroimaging, where we want to identify neural activities using MEG sensor measurements outside the brain. The problem is ill-posed since the observed magnetic field could result from an infinite number of possible neuronal sources. In this paper, we propose a family of methods for coefficient screening by using sensor covariance thresholding and shrinkage. The new methods assume that the structure of sensor measurements can be modelled by a set of non-orthogonal covariance components. We develop an asymptotic theory for identifying non-zero coefficients estimators. We also derive the lower and upper bounds for the mean screening errors of the proposed methods under certain conditions. The new theory is further illustrated by simulations and a real data analysis.

Classification Analysis of Big Image Data

Nan Lin, Junhai Jiang, Shicheng Guo, Xiao Yu, Long Ma and Momiao Xiong*

Division of Biostatistics, The University of Texas Health Science Center at Houston

Due to advances in sensors, growing large and complex medical images provides invaluable information for holistic discovery of the genetic and epigenetic structure of disease and has the potential to enhance diagnosis of disease, prediction of clinical outcomes, characterization of disease progression, management of health care and development of treatments, but also pose great methodological and computational challenges. An enormous amount of increasingly larger, more complex and more diverse demand developing unified frameworks and novel statistical methods for cluster and classification analysis of medical image data, which will provide low-cost and powerful tools for early detection and efficient management of complex diseases such as cancers, mental disorders, vascular diseases. The medical images have the ability to visualize the pathology change in the cellular or even the molecular level or anatomical changes in tissues and organs. However, the medical images for the same type of disease from different individuals might be quite similar. As a result, it is a big challenge to extract the key information from a large amount of medical images for early detection of the complex diseases and the prediction of the drug response. To address this issue, we extend one dimensional functional principal component analysis to the two

dimensional functional principle component analysis (2DFPCA). To reduce high dimensional image data to low dimensional space, we develop novel space sufficient dimension reduction methods to select variables. The proposed methods are applied to 250 liver cancer histology image data (99 tumor tissues and 151 normal tissues) and 176 ovarian cancer histology images with the drug response status from TCGA database. For the liver cancer dataset, we can reach almost 84%, 79.8% and 86.8% classification accuracy, sensitivity and specificity, respectively. For the ovarian cancer drug response dataset, classification accuracy, sensitivity and specificity are 80.1%, 85.8% and 71.4, respectively.

Recent Developments of Iterative Monte Carlo Methods for Big Data Analysis

Faming Liang and Chuanhai Liu

Taxes A&M University and Purdue University

Iterative Monte Carlo methods, such as MCMC, stochastic approximation, and EM, have proven to be very powerful tools for statistical data analysis. However, their computer-intensive nature, which typically require a large number of iterations and a complete scan of the full dataset for each iteration, precludes their use for big data analysis. We will provide an overview of the recent developments of iterative Monte Carlo methods for big data analysis. The portion by Liang will focus on the developments of the MCMC and stochastic approximation methods, and that by Liu will focus on the developments of the EM method.

A Partial Review of Software for Big Data Statistics

Jun Yan

Department of Statistics, University of Connecticut

Big data brings challenges to even simple statistical analysis because of the barriers in computer memory and computing time. The computer memory barrier is usually handled by a database connection that extracts data in chunks for processing. The computing time barrier is handled by parallel computing, often accelerated by graphical processing units. In this partial review, we summarize the open source R packages that break the computer memory limit such as biglm and bigmemory, as well as the academic version of the commercial Revolution R, and R packages that support parallel computing. Products from commercial software will also be sketched for completeness. Joint work with Ming-Hui Chen, Elizabeth Schifano, Chun Wang, and Jing Wu of University of Connecticut.

Wednesday, February 12

Numerical Aspects of Bayesian Model Selection in High and Ultrahigh-dimensional Settings

Valen Johnson

Department of Statistics, Texas A&M University

In this talk, I examine the empirical convergence properties of a Bayesian model selection procedure based on a non-local prior density in ultrahigh-dimensional settings. The performance of the model selection procedure is also compared to popular penalized likelihood methods. Coupling diagnostics are used to bound the total variation distance between iterates in an Markov chain Monte Carlo (MCMC) algorithm and the posterior distribution on the model space. In several simulation scenarios in which the number of observations exceeds 100, rapid convergence and high accuracy of the Bayesian procedure is demonstrated. Conversely, the coupling diagnostics are successful in diagnosing lack of convergence in several scenarios for which the number of observations is less than 100. The accuracy of the Bayesian model selection procedure in identifying high probability models is shown to be comparable to commonly used penalized likelihood methods, including extensions of smoothly clipped absolute deviations (SCAD) and least absolute shrinkage and selection operator (LASSO) procedures.

Quantile Regression in Variable Screening

Lingsong Kong

Department of Statistics, Purdue University

We introduce a quantile regression framework for linear and nonlinear variable screening with high-dimensional heterogeneous data. Motivated by success of various variable screening methods, especially the quantile-adaptive framework, we propose to combine the information from different quantile levels to provide more efficient variable screening procedure. In particular, there are two ways to do so: one is to simply take (weighted) average across different levels of quantile regression; the other one is to use (weighted) composite quantile regression. Asymptotically, these two approaches are equivalent in terms of efficiency. Numerical studies confirm the fine performance of the proposed method for various linear and nonlinear models. Joint work with Qian Shi.

What Is Beyond Sparse Coding?

Ying Nian Wu

Department of Statistics, University of California at Los Angeles

Many types of data such as natural images admit sparse representations by redundant dictionaries of basis functions (or regressors), and these dictionaries can either be designed or learned from training data. However, it is still unclear how to go beyond sparsity and continue to learn structures behind the sparse representations. In this talk, I shall review some recent progresses and the major issues and difficulties that need to be addressed. I shall also present our own recent work that seeks to learn dictionaries of compositional patterns in the sparse representations. Based on joint work with Jianwen Xie, Wenze Hu and Song-Chun Zhu.

Functional Regression and Image Regression

Xiao Wang

Department of Statistics, Purdue University

The first part of this talk considers prediction and testing with functional predictors in the framework of functional linear model and reproducing kernel Hilbert space. The lower bounds for both the minimax prediction and the minimax separation distance of the slope function are derived. It is shown that the optimal rates are determined jointly by the reproducing kernel and the covariance kernel, but the rates are different for prediction and testing. An easily implementable roughness regularization predictor is shown to attain the optimal rate of convergence for prediction. Further, a generalized likelihood ratio test attains the optimal rate of convergence for testing. The second part considers prediction with image predictors in the framework of functional linear model and bounded total variation space. We assume that the slope function belongs to the space of bounded total variation. We further assume that the slope function can be generated by a finite number of wavelet basis functions, but we do not specify how many and which basis functions enter the model. Efficient algorithm is developed to estimate the slope function. We show that our procedures are consistent in the sense of prediction and selecting the model correctly.

Sentiment Analysis

Xiaotong Shen

School of Statistics, University of Minnesota

Sentiment analysis identifies the relevant content as well as determines and understands opinions, from documents or texts, towards a specific event of interest. In this presentation, I will discuss large margin methods for ordinal classification involving word predictors, where imprecise information is available for prediction regarding linguistic relations among predictors, expressed in terms of a directed graph. Then the methods will be used for sentiment analysis, where sentiment function representations of words are derived, on which the imprecise predictor relations are integrated as linear relational constraints over sentiment function coefficients. Computational and theoretical aspects will be discussed, in addition to an application to opinion survey.

A Bayesian Approach to Subgroup Identification

James O. Berger, Xiaojing Wang*, and Lei Shen

Department of Statistical Science, Duke University
Department of Statistics, University of Connecticut
Eli Lilly and Company

The paper discusses subgroup identification, the goal of which is to determine the heterogeneity of treatment effects across subpopulations. Searching for differences among subgroups is challenging because it is inherently a multiple testing problem with the complication that test statistics for subgroups are typically highly dependent, making simple multiplicity corrections such as the Bonferroni correction too conservative. In this paper, a Bayesian approach to identify subgroup effects is proposed, with a scheme for assigning prior probabilities to possible subgroup effects that accounts for multiplicity and yet allows for (pre-experimental) preference to specific subgroups. The analysis utilizes a new Bayesian model selection methodology and, as a byproduct, produces individual probabilities of treatment effect that could be of use in personalized medicine. The analysis is illustrated on an example involving subgroup analysis of biomarker effects on treatments.

Detection of Tumor Driver Genes Using a Fully Integrated Bayesian Approach

Guanghua Xiao

Department of Clinical Science, UT Southwestern Medical Center

DNA copy number alterations (CNAs), including amplifications and deletions, can result in significant changes in gene expression, and are closely related to the development and progression of many diseases, especially cancer. For example, CNA-associated expression changes in certain genes (called tumor driver genes) can alter the expression levels of many downstream genes through transcription regulation, and cause cancer. Identification of such tumor driver genes leads to discovery of novel therapeutic targets for personalized treatment of cancers. Several approaches have been developed for this purpose by using both copy number and gene expression data.

In this study, we propose a Bayesian approach to identify tumor driver genes, in which the copy number and gene expression data are modeled together, and the dependency between the two data types is modeled through conditional probabilities. The proposed joint modeling approach can identify CNA and differentially expressed (DE) genes simultaneously, leading to improved detection of tumor driver genes and comprehensive understanding of underlying biological processes. The proposed method was evaluated in simulation studies, and then applied to a head and neck squamous cell carcinoma (HNSCC) dataset. Both simulation studies and data application show that the joint modeling approach can significantly improve the performance in identifying tumor driver genes, when compared to other existing approaches.

Online Updating of Statistical Inference in the Big Data Setting

Elizabeth Schifano

Department of Statistics, University of Connecticut

We present statistical methods for big data arising from online analytical processing, where large amounts of data arrive in streams and require fast analysis without storage/access to the historical data. In particular, we develop iterative estimating algorithms and statistical

inferences for linear models and estimating equations that update as new data arrive. The online updating framework in the linear model setting introduces predictive residuals that can be used to test the goodness-of-fit of the hypothesized model. In simulation studies, our approach compares favorably with competing approaches in terms of timing and accuracy. Joint work with Ming-Hui Chen, Jun Yan, Chun Wang, Jing Wu (Department of Statistics, University of Connecticut)

Statistical Aggregation in Massive Data Environment

Nan Lin

Department of Mathematics, Washington University in St. Louis

Due to their size and complexity, massive data sets bring many computational challenges for statistical analysis, such as overcoming the memory limitation and improving computational efficiency of traditional statistical methods. In this talk, I will discuss the statistical aggregation strategy to conquer such challenges posed by massive data sets. Statistical aggregation partitions the entire data set into smaller subsets, compresses each subset into certain low-dimensional summary statistics and aggregates the summary statistics to approximate the desired computation based on the entire data. Results from statistical aggregation are required to be asymptotically equivalent. Statistical aggregation is particularly useful to support sophisticated statistical analyses for online analytical processing in data cubes. We will detail its application to two large families of statistical methods, estimating equation estimation and U-statistics.

Thursday, February 13

Detecting Genetic Association Signals Leveraging Network Information

Hongyu Zhao

Yale School of Public Health

Although Genome Wide Association Studies (GWAS) have identified many sceptibility loci for common diseases, these loci only explain a small portion of heritability. It is challenging to identify the remaining disease loci because their association signals are likely weak and difficult to identify among millions of candidates. One potentially useful direction to increase statistical power is to incorporate pathway and functional genomics information to prioritize GWAS signals. In this talk, we first describe a method to utilize network information to prioritize disease genes based on the 3guilt by association2 principle, in which networks are treated as static, and disease associated genes are assumed to locate closer with each other than random pairs in the network. We then introduce a novel 3guilt by rewiring2 principle that postulates that disease genes more likely undergo rewiring in disease patients, whereas most of the network is unaffected in disease condition. A Markov random field framework was used for both methods to integrate network information to prioritize genes. Applications in Crohn1sdisease and Parkinson1s disease show that these methods lead to more replicable and biologically meaningful results. This is joint work with Min Chen, Lin Hou, Clarence Zhang, and Judy Cho.

Biocuration in the Era of Big Data

Zhang Zhang

CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

With the rapid advancements in high-throughput sequencing technologies, biology enters the era of big data. Many databases developed for managing biological data are traditionally based on expert curation, viz., conducted manually by dedicated experts. However, with the burgeoning volume of biological data and increasingly diverse densely informative published literatures, expert curation becomes more and more laborious and time consuming, increasingly lagging behind knowledge creation, or worse, not being done at all in fields where insufficient funds can be allocated to curation. Although traditionally expert-curated databases

have proven important for biological studies, they are struggling with the flood of knowledge and accordingly requiring a large number of people getting involved in curation, viz., community curation-exploiting the whole power of the scientific community for knowledge integration.

A case in point that harnesses community intelligence in knowledge integration is Wikipedia. Wikipedia is an online encyclopedia, allows anyone to create/edit any content and features collaborative knowledge curation, up-to-date content, huge coverage, and low cost for maintenance. Despite fears that the openness of editorial capacity could lead to incorporation of significant flawed content, it is reported that Wikipedia rivals the traditional encyclopedia in accuracy. Due to the extraordinary success of Wikipedia, it has been advocated that biological databases go wiki. As a consequence, more than a dozen biological wikis (bio-wiki) have been constructed to call on community intelligence in knowledge curation. To date, however, there is no community-curated resource for rice, as rice is the most important staple food feeding a large part of the world population and building expert-curated rice reference genomes with comprehensive and accurate annotations remains a formidable challenge. Moreover, one of the major limitations in bio-wikis is insufficient participation from the scientific community, which is intrinsically because of lack of explicit authorship and thus no credit for community-curated contributions.

To increase community curation in bio-wikis, we developed AuthorReward (Bioinformatics 2013), to reward community-curated efforts by contribution quantification and explicit authorship. AuthorReward quantifies researchers' contributions by properly factoring both edit quantity and quality and yields automated explicit authorship according to their quantitative contributions. Author Reward provides bio-wikis with an authorship metric, helpful to increase community participation in bio-wikis and to achieve community curation of massive biological knowledge. We also developed RiceWiki (http://ricewiki.big.ac.cn; Nucleic Acids Research 2014), a wiki-based, publicly editable, and open-content platform for community curation of rice genes. To test the functionality of AuthorReward, we installed it in RiceWiki. A live demo is the rice semi-dwarfing gene (sd1), which was collaboratively curated by 9 researchers, providing 89 versions as of August 1, 2013. As testified in RiceWiki, AuthorReward is capable of yielding sensible quantitative contributions and providing automated explicit authorship, consistent well with perceptions of all participated contributors. Additionally, due to significant importance of rice, RiceWiki serves as a critical community-curated knowledgebase for the rice research community. Considering the growing volume of rice-related data and contrastingly the small number of expert curators working on rice, RiceWiki bears the potential to make it possible to build a rice encyclopedia by and for the scientific community, which harnesses collective intelligence for collaborative knowledge curation, covers all aspects of biological knowledge, and keeps evolving with novel knowledge.

Poly(A) Motif Prediction Using Spectral Latent Features from Human DNA Sequences

Xin Gao

King Abdullah University of Science and Technology
Kingdom of Saudi Arabia

Polyadenylation is the addition of a poly(A) tail to an RNA molecule. Identifying DNA sequence motifs that signal the addition of poly(A) tails is essential to improved genome annotation and better understanding of the regulatory mechanisms and stability of mRNA. Existing poly(A) motif predictors demonstrate that information extracted from the surrounding nucleotide sequences of candidate poly(A) motifs can differentiate true motifs from the false ones to a great extent. A variety of sophisticated features has been explored, including sequential, structural, statistical, thermodynamic and evolutionary properties. However, most of these methods involve extensive manual feature engineering, which can be time-consuming and can require in-depth domain knowledge.

We propose a novel machine learning method for poly(A) motif prediction by marrying generative learning (hidden Markov models) and discriminative learning (support vector machines). Generative learning provides a rich palette on which the uncertainty and diversity of sequence information can be handled, while discriminative learning allows the performance of the classification task to be directly optimized. Here, we employed hidden Markov models for fitting the DNA sequence dynamics, and developed an efficient spectral algorithm for extracting latent variable information from these models. These spectral latent features were then fed into support vector machines to fine tune the classification performance. We evaluated our proposed method on a comprehensive human poly(A) dataset that consists of 14,740 samples from 12 of the most abundant variants of human poly(A) motifs. Compared with one of previous state-of-art methods in the literature (the random forest model with expert-crafted features), our method reduces the average error rate, false negative rate and false positive rate by 26%, 15% and 35%, respectively. Meanwhile, our method made about 30% fewer error predictions relative to the other string kernels. Furthermore, our method can be used to visualize the importance of oligomers and positions in predicting poly(A) motifs, from which we can observe a number of characteristics in the surrounding regions of true and false motifs that have not been reported before.

Statistical Inference for Massive Distributed Spatial Data Using Low-rank Models

Matthias Katzfuss

Department of Statistics, Texas A&M University

Due to rapid data growth, it is increasingly becoming infeasible to move massive datasets, and statistical analyses have to be carried out where the data reside. If several massive datasets stored in separate physical locations are all relevant to a given problem, the challenge is to obtain valid inference based on all data without moving the datasets. This distributed data problem frequently arises in the geophysical and environmental sciences, for example when a spatial process of interest is measured by several satellite instruments. We show that for the widely used class of spatial low-rank models, which contain a component that can be written as a linear combination of spatial basis functions, computationally feasible spatial inference and prediction for massive distributed data can be carried out exactly and in parallel. The required number of floating-point operations is linear in the number of data points, while the required amount of communication does not depend on the data sizes at all.

After discussing several extensions and special cases of this result, we apply our methodology to carry out spatio-temporal filtering inference on total precipitable water measured by three different sensor systems.

Annealing Between Distributions by Averaging Moments

Ruslan Salakhutdinov

Department of Computer Science and Department of Statistical Sciences
University of Toronto

Many powerful Monte Carlo techniques for estimating partition functions, such as annealed importance sampling (AIS), are based on sampling from a sequence of intermediate distributions which interpolate between a tractable initial distribution and the intractable target distribution. The near-universal practice is to use geometric averages of the initial and target distributions, but alternative paths can perform substantially better. We present a novel sequence of intermediate distributions for exponential families defined by averaging the moments of the initial and target distributions. We analyze the asymptotic performance of both the geometric and moment averages paths and derive an asymptotically optimal piecewise linear schedule. AIS with moment averaging performs well empirically at estimating partition functions of restricted Boltzmann machines (RBMs), which form the building blocks of many

deep learning models, including Deep Belief Networks and Deep Boltzmann Machines. Joint work with Roger Grosse and Chris Maddison

References:

Annealing between Distributions by Averaging Moments. Roger Grosse, Chris Maddison, and Ruslan Salakhutdinov. In Neural Information Processing Systems (NIPS 27) www.cs.toronto.edu/~rsalakhu/papers/nips2013_moment.pdf

MCMC for Non-Linear State Space Models Using Ensembles of Latent Sequences

Alex Shestopaloff

Department of Statistical Sciences, University of Toronto

Non-linear state space models are a widely-used class of models for biological, economic, and physical processes. Fitting these models to observed data is a difficult inference problem that has no straightforward solution. We take a Bayesian approach to the inference of unknown parameters of a non-linear state model; this, in turn, requires the availability of efficient Markov Chain Monte Carlo (MCMC) sampling methods for the latent (hidden) variables and model parameters. Using the ensemble technique of Neal (2010) and the embedded HMM technique of Neal (2003), we introduce a new Markov Chain Monte Carlo method for non-linear state space models. The key idea is to perform parameter updates conditional on an enormously large ensemble of latent sequences, as opposed to a single sequence, as with existing methods. We look at the performance of this ensemble method when doing Bayesian inference in the Ricker model of population dynamics. We show that for this problem, the ensemble method is vastly more efficient than a simple Metropolis method, as well as 1.9 to 12.0 times more efficient than a single-sequence embedded HMM method, when all methods are tuned appropriately. We also introduce a way of speeding up the ensemble method by performing partial backward passes to discard poor proposals at low computational cost, resulting in a final efficiency gain of 3.4 to 20.4 times over the single-sequence method. This research has been done jointly with Radford M. Neal.

The Screening and Ranking Algorithm for Change-Points Detection in Multiple Samples

Xiaoyi Min

Yale School of Public Health

DNA copy number variation (CNV) is a form of genomic structural variation that may affect human diseases. Identification of the CNVs shared by many people in the population as well as determining the carriers of these CNVs is essential for understanding the role of CNV in disease association studies. For detecting CNVs in single samples, a Screening and Ranking Algorithm (SaRa) was previously proposed, which was shown to be superior over other commonly used algorithms and have a sure coverage property. We extend SaRa to address the problem of common CNV detection in multiple samples. In particular, we propose an adaptive Fisher's method for combining the screening statistics across samples. The proposed multi-sample SaRa method inherits the computational and practical benefits of single sample SaRa in CNV detection. We also characterize the theoretical properties of this method and demonstrate its performance in extensive numerical analyses.

Variance Estimation for High-Dimensional Linear Models

Lee Dicker

Department of Statistics, Rutgers University

The residual variance and the proportion of explained variation are important quantities in many statistical models and model fitting procedures. They play an important role in regression diagnostics, model selection procedures, and in determining the performance limits in many problems. Recently, methods for estimating these and other related summary statistics in high-dimensional linear models have received significant attention. In this talk, we discuss some of the various approaches to estimating these quantities (e.g., residual sum-of-squares-based estimators, the method-of-moments) and the conditions required to ensure reliable performance (sparsity, conditions on the predictor covariance matrix). Efficiency will also be discussed, along with new estimators that are closely related to ridge regression. We will consider an application related to estimating heritability, an important concept in genetics.

D&R for Large Complex Data: Likelihood Modelling for Logistic Regression

Philip Gautier*, William S. Cleveland, and Chuanhai Liu

Department of Statistics, Purdue University

Divide and recombine (D&R) is a statistical framework for the aAlex Shestopaloffnalysis of large, complex data. The data are divided into subsets. Numeric and visualization methods, which collectively are analytic methods, are applied to each subset. For each analytic method, the outputs of the application of the method to the subsets are recombined. D&R computation for the application of an analytic method is embarrassingly parallel: the subset computations are independent and do not communicate with one another. Here we study D&R methods for likelihood-based model fitting. We introduce a notion of likelihood analysis and modelling. We divide the data and fit a likelihood model on each subset. The fitted model is characterized by a set of parameters much smaller than the subset data size, but retains as much information as possible about the true subset likelihood. Analysis of subset likelihoods and their fitted models consists of visualizations on an appropriate scale and region. These visualizations allow the analyst to verify the choice and fit of the model. The fitted models are recombined across subsets to form a model of the all-data likelihood, which we maximize to obtain an estimate for the all-data MLE. We present simulation results demonstrating the performance of our method compared with the all-data MLE for the case of logistic regression.