The city of Sanya

welcomes you!

Big Data: Opportunities, Challenges and Innovations

Dec 27-30, 2014

Sanya, Hainan

Keynote Session 1

- ***Chair: Rong Chen***

Big Data in Idea Space and in Regulatory Space

Richard B. Freeman

Harvard University, Cambridge, MA, USA
Email: rbfreeman@gmail.com

My presentation has two parts: 1) Analysis of big data in the study of the production and use of scientific ideas, among scientists and institutions, with particular attention to linkages between different networks – co-authors, citations, scientific instruments – and their relation to funding. This represents the work of my NBER group on science and engineering; 2) Analysis of governmental administrative data to improve regulation of firms, which involves linking multiple data sets from diverse agencies to discover the presumptively small set of firms that cause many problems to society, and for analysis of the pattern of inequality and financial risks. In both parts I will present hypotheses that we will be exploring in the next year or two, and potential use of graph theory to test these hypotheses.

Session 1: Economics & Business

- *Chair: Steve Z. Qin*

Modeling and Forecasting of Chinese CPI

Song Xi Chen

Guanghua School of Management, Peking University, Beijing, China
Email: csx@gsm.pku.edu.cn

This paper starts with a detailed analysis on the micro-dynamic structure of the Consumer price index (CPI) series from China, revealing the seasonality and the effect of the Spring festival on the price index. We then show that the CPI from China is quite predictable in terms of the simulated out-of-the sample prediction error.

A Statistical Model for Social Network Labeling

Hansheng Wang

Guanghua School of Management, Peking University, Beijing, China
Email: hansheng@gsm.pku.edu.cn

We consider a social network from which one observes not only network structure (i.e., nodes and edges) but also a set of labels (or tags, keywords) for each node (or user). These labels are self-created and closely related to the user's career status, life style, personal interests, and many others. Thus, they are of great interest for online marketing. To model their joint behavior with network structure, a complete data model is developed. The model is based on the classical p_1 model but allows the reciprocation parameter to be label-dependent. By focusing on connected pairs only, the complete data model can be generalized into a conditional model. Compared with the complete data model, the conditional model specifies only the conditional likelihood for the connected pairs. As a result, it suffers less risk from model mis-specification. Furthermore, because the conditional model involves connected pairs only, the computational cost is much lower. The resulting estimator is consistent and asymptotically normal. Depending on the network sparsity level, the convergence rate could be different. To demonstrate its finite sample performance, numerical studies (based on both simulated and real datasets) are presented.

The Sales Effects of Having Friends: A Two-Phase Identification Method with Observational Data

Junni L. Zhang

[1]Guanghua School of Management, Peking University, Beijing, China
Email: zjn@gsm.pku.edu.cn

Despite the theoretical importance of social influence, little research has empirically examined the causal sales impact of having friends (vs. not) with observational data. To quantify this impact,we propose an integrated two-phase causality inference model. The first phase uses propensity score matching to assure randomized treatment and address sample selection endogeneity due to observed variables. The second phase leverages principal stratification to assure a common set of treated and control consumers and address both data truncation bias and unobservable variables in the post-treatment period. This common set has well-defined consumer purchases both when having friends and when not having friends. We apply our model to a massive multiplayer online role playing game. We find that without fully balancing the statistical properties of covariates among heterogeneous consumers, researchers risk dramatically over-estimating the sales impact of having friends, with large spurious results. Interestingly, without modeling sales outcome data truncation or unobservables, researchers with only propensity score matching mightseriously under-estimate the true causal sales impact of having friends. This two-phase causality identification method is novel in the literature and useful to managers with an abundance of observational data.

Keynote Session 2

- ❖ *Chair: Dongchu Sun*

Equivalent Partial Correlation Selection for High Dimensional Gaussian Graphical Models

Faming Liang

Department of Biostatistics, University of Florida, Gainesville, FL, USA
Email: faliang@ufl.edu

Gaussian graphical models (GGMs) are frequently used to explore networks, such as gene regulatory networks, among a set of variables. Under the classical theory of GGMs, the construction of Gaussian graphical networks amounts to finding the pairs of variables with nonzero partial correlation coefficients. However, this is infeasible for high dimensional problems for which the number of variables is larger than the sample size. We propose a new measure of partial correlation coefficient, which is evaluated with a reduced conditional set and thus feasible for high dimensional problems. Under the Markov property and adjacency faithfulness conditions, the new measure of partial correlation coefficient is equivalent to the true partial correlation coefficient in construction of Gaussian graphical networks. Based on the new measure of partial correlation coefficient, we propose a multiple hypothesis test-based method for construction of Gaussian graphical networks. Further, we establish the consistency of the proposed method under mild conditions. The proposed method outperforms the existing methods, such as the PC, graphical Lasso, nodewise regression, and qp-average methods, especially for the problems for which a large number of indirect associations are present. The proposed method has a computational complexity of nearly $O(p^2)$, and is flexible in data integration, network comparison, and covariate adjustment.

Session 2: Statistical Theory & Methodology I

❖ *Chair: Wenxuan Zhong*

Jun S. Liu

Harvard University
Email: jlliu1600@gmail.com

TBD.

Bayesian Analysis of Multivariate Smoothing Splines

Dongchu Sun[1], Zhuoqiong He[2]

[1]University of Missouri and East China Normal University
[2]University of Missouri
Email: sund@missouri.edu

A general version of multivariate smoothing spline with correlated errors and correlated curves is introduced. A suitable symmetric smoothing parameter matrix is introduced, and practical priors are developed for the unknown covariance matrix of the errors and the smoothing parameter matrix. An efficient algorithm for computing the multivariate smoothing spline is derived, which leads to an efficient Markov chain Monte Carlo method for Bayesian computation. Key to the computation is a natural decomposition of the estimated curves into components intrinsic to the problem that extend the notion of principal components. These intrinsic principal curves are useful both for computing and for interpreting the data. The methods are applied to the multivariate yield curves of Chinese Government Bonds and Bank Bonds.

Measuring the reproducibility of high-throughput experiments and covariate effects of influencing factors

Qunhua Li

Dept. of Statistics, Pennsylvania State University, USA
Email: qunhua.li@psu.edu

Reproducibility is essential to scientific discoveries from the high-throughput experiments. The outcome of such experiments is affected by many operational factors in the experimental and

data-analytical procedures. Understanding how these factors affect the reproducibility of the outcome is critical for establishing workflows that produce replicable discoveries.

In this talk, I will present two pieces of our work on the reproducibility of high-throughput experiment. The first part assesses the reproducibility of findings from replicate high-throughput experiments using a copula mixture model. By jointly modeling the significance and consistency of the signals across replicates, this model computes a reproducibility index for each finding, in a fashion analogous to FDR. It allows findings to be ranked and selected by their reproducibility. It is adopted by ENCODE and modENCODE consortia for processing the production ChIP-seq data. The second part is a regression framework to assess the covariate effect of operational factors on the reproducibility of findings from high-throughput experiments. In contrast to the existing graphical approaches, our regression framework allows one to succinctly characterize the simultaneous and independent effects of covariates on reproducibility and to compare reproducibility while controlling for potential confounding variables.

We illustrate the usefulness of our methods using ChIP-seq and microarray studies.

Phase Transition and Regularized Bootstrap in Large-scale *t*-tests with False Discovery Rate Control

Weidong Liu

Shanghai Jiao Tong University, Shanghai, China
Email: weidongl@sjtu.edu.cn

Applying the Benjamini and Hochberg (B-H) method to multiple Student's *t*-tests is a popular technique for gene selection in microarray data analysis. Given the non-normality of the population, the true p-values of the hypothesis tests are typically unknown. Hence, it is common to use the standard normal distribution $N(0,1)$, Student's t distribution tn−1 or the bootstrap method to estimate the p-values. In this paper, we prove that when the population has the finite 4-th moment and the dimension m and the sample size n satisfy $\log m=o(n^{1/3})$, the B-H method controls the false discovery rate (FDR) and the false discovery proportion (FDP) at a given level α asymptotically with p-values estimated from $N(0, 1)$ or t_{n-1} distribution. However, a phase transition phenomenon occurs when $\log m \geq c_0 n^{1/3}$. In this case, the FDR and the FDP of the B-H method may be larger than α or even con- verge to one. In contrast, the bootstrap calibration is accurate for $\log m = o(n^{1/2})$ as long as the underlying distribution has the sub-Gaussian tails. However, such a light-tailed condition cannot generally be weakened. The simulation study shows that the bootstrap calibration is very conservative for the heavy tailed distributions. To solve this problem, a regularized bootstrap correction is proposed and is shown to be robust to the tails of the distributions. The simulation study shows that the regularized bootstrap method performs better than its usual counterpart.

Session 4: Life Science I

- ***Chair: Ke Deng***

Opportunities and Challenges in Analyzing Big Metagenome Data

Xuegong Zhang[1,2]

[1]Bioinformatics Div, TNLIST and Department of Automation, Tsinghua University, Beijing, China
[2]School of Life Sciences, Tsinghua University, Beijing, China
Email: zhangxg@tsinghua.edu.cn

Current biological and medical researches characterized with the heavy involvement of various types of omic data provide some representative examples of big-data-driven scientific research. Among other properties, they possess an unique feature that the data of each single instance in the study are getting overwhelmingly huge, but the number of instances that can be investigated in most studies remain limited, and the data usually require multiple challenging and usually error-prone processing steps before any biological sense can be made. Mategenomics data are a typical example such big data. It has been recently found that the mixture of microbial habitants (the microbiome) inside and on the human body can play important roles in human health. The number of microbial cells and genes are orders higher than that of human cells and human genes. Metagenome data are the DNA sequencing of mixed genomes of unknown microbiomes. They can be viewed as the random sampling of short sequencing reads from mixtures of unknown number of bacteria and/or archaea genomes. The data can easily reach hundreds of gigabytes for a single individual sample, and up to dozens of terabytes for a moderate study. An important use of such data is the mining of features of microbial taxonomic compositions, of microbial genes and pathways, and of sequence signatures that can be used to compare and characterize individuals hosting the microbiomes or that are associated with phenotypical features of the individuals. Similar studies can also be done on evironmental samples such as microbial mixtures in water and soils. Challenges exist in both the low-level processing and quantitative inferences of the raw metagenome sequencing data, and in the high-level analysis of the taxonomic and functional information derived from the raw data, and in developing new ways to analyze the data that do not reply on known taxonomic or functional annotations. In this talk, we'll try to give an overview of some major data analysis challenges in this field, with an emphasis on high-level data analysis tasks. We will also share our on-going practices on the application of unsupervised and supervised machine learning methods on metagenomic data with some preliminary results on real data.

A Dynamic Directional Model for Effective Brain Connectivity Using Electrocorticographic (ECoG) Time Series

Tingting Zhang

Department of Statistics, University of Virginia, USA
Email: tz3b@virginia.edu

We introduce a dynamic directional model (DDM) for studying brain effective connectivity based on intracranial electrocorticographic (ECoG) time series. The DDM consists of two parts: a set of differential equations describing neuronal activity of brain components (state equations), and observation equations linking the underlying neuronal states to observed data. When applied to functional MRI or EEG data, DDMs usually have complex formulations and thus can accommodate only a few regions, due to limitations in spatial resolution and/or temporal resolution of these imaging modalities. In contrast, we formulate our model in the context of ECoG data. The combined high temporal and spatial resolution of ECoG data result in a much simpler DDM, allowing investigation of complex connections between many regions. To identify functionally segregated sub-networks, a form of biologically economical brain networks, we propose the Potts model for the DDM parameters. The neuronal states of brain components are represented by cubic spline bases and the parameters are estimated by minimizing a log-likelihood criterion that combines the state and observation equations. The Potts model is converted to the Potts penalty in the penalized regression approach to achieve sparsity in parameter estimation, for which a fast iterative algorithm is developed. The methods are applied to an auditory ECoG dataset.

A metagenomic study on the inhalable microorganisms in Beijing's PM pollutants

Ting F. Zhu

School of Life Sciences, Tsinghua University, Beijing, China
Email: tzhu@biomed.tsinghua.edu.cn

Particulate matter (PM) air pollution poses a formidable public health threat to the city of Beijing. Among the various hazards of PM pollutants, microorganisms in $PM_{2.5}$ and PM_{10} are thought to be responsible for various allergies and for the spread of respiratory diseases_ENREF_1_6. While the physical and chemical properties of PM pollutants have been extensively studied, much less is known about the inhalable microorganisms. Most existing data on airborne microbial communities using 16S or 18S rRNA gene sequencing to categorize bacteria or fungi into the family or genus levels do not provide information on their allergenic and pathogenic potentials. Since most existing data on airborne bacteria communities using 16S rRNA gene sequencing to categorize them into the family or genus level do not provide information on their allergenic and pathogenic potential, we employed metagenomic tools to sequence the inhalable airborne

microorganisms carried by Beijing's $PM_{2.5}$ and PM_{10} pollutants during January 8-14, 2013. By aligning to a cohort of 2637 non-redundant NCBI complete genomes, and by using the Metagenomic Phylogenetic Analysis (MetaPhlAn) toolbox and the Metagenomics RAST (MG-RAST) server for taxonomic assignments, we identified airborne microbes including bacteria, archaea, fungi, and viruses at the species level. Our results suggested that the majority of the inhalable microorganisms were soil-associated and non-pathogenic to human. Yet several respiratory allergens and pathogens (e.g., *Streptococcus pneumoniae*, *Aspergillus fumigatus*, and human adenovirus C) were identified in Beijing's PM pollutants, and their relative abundance appeared to have increased with progressively worsened pollution levels. The proportion of airborne bacteria from terrestrial-related sources was found to be higher in Beijing than those previously identified from other places, and the fraction of fecal-associated bacteria appeared to have increased with deteriorated PM pollution levels. The existence of respiratory allergens and pathogens in Beijing's PM pollutants, especially with their increased relative abundance with pollution levels, may pose threats to the susceptible population, causing allergies and leading to the spread of respiratory diseases. Our findings may serve as an important reference in respiratory medicine and environmental science for the efforts to reduce PM pollution and to prevent respiratory diseases and allergies.

Omicseq: a genomics BigData search engine and knowledge discovery system

Zhaohui Qin

Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA
Email: zhaohui.qin@emory.edu

There are millions of publicly available genomics datasets with more being generated daily at an accelerated pace. However, the vast majority of these datasets remain underutilized after their initial publication due to the difficulty of processing and analyzing them. Consequently there is much more information in these datasets than is being reported. Another problem is how to identify datasets that a given researcher is most interested in, which is analogous to identifying the most relevant webpages for a given domain of address. To address these problems, we started the Omicseq project which has two goals: 1. Build an IT infrastructure to better organize publicly available genomics data. 2. Develop an innovative biomedical data browser to efficiently browse through these datasets to identify novel biological knowledge and insights. We will discuss our current efforts and strategies for future development.

❖ ***Chair: Jun S. Liu***

Normalizing large, heterogeneous datasets

Terence Paul Speed

Bioinformatics Division, Walter & Eliza Hall Institute of Medical Research, Australia
Email: terry@wehi.edu.au

In my field, large datasets are usually heterogeneous, obtained by aggregating smaller datasets. For example, the data studied by Kang *et al* (Spatio-temporal transcriptome of the human brain, **Nature** 2011) came from 1,340 samples of post- mortem brain tissue from 57 developing and adult brains taken from 16 regions of the brain at 15 time periods. Each sample gives one microarray gene expression dataset. Large-scale tumour studies, of which there are now many, can exhibit similar heterogeneity. For most analyses of such large datasets, some normalization or adjustment will be necessary, where these terms loosely mean dealing with the heterogeneity, whether that is due to technical artefacts of the measurement process, features of the samples not of interest in the current analysis, or something else. The nature of the heterogeneity may be known in part, or quite unknown.

In my talk, I'll discuss a class of methods of dealing with heterogeneity without assuming its cause is known which make use of what we term *negative controls*. I'll begin by describing the class of linear models we use in our approach, and illustrate it in four contexts: that of identifying differentially expressed genes, classifying, clustering, and correlating. The methods apply to a wide variety of kinds of data, but for this talk, my illustrations will all come from microarray gene expression studies.

Session 4: Statistical Theory & Methodology II

❖ ***Chair: Tingting Zhang***

Oracle-efficient confidence envelopes for covariance functions in dense functional data

Lijian Yang

Center for Advanced Statistics and Econometrics Research, Soochow University, Suzhou, China
Email: yanglijian@suda.edu.cn

We consider nonparametric estimation of the covariance function for dense functional data using computationally efficient tensor product B-splines. We develop both local and global asymptotic distributions for the proposed estimator, and show that our estimator is as efficient as an “oracle” estimator where the true mean function is known. Simultaneous confidence envelopes are developed based on asymptotic theory to quantify the variability in the covariance estimator and to make global inferences on the true covariance. Monte Carlo simulation experiments provide strong evidence that corroborates the asymptotic theory. Two real data examples on the near infrared spectroscopy data and speech recognition data are also provided to illustrate the proposed method.

Population Size Estimation with Covariate Values Missing Non-ignorable

Liping Liu

School of Mathematical Sciences, Peking University, Beijing, China
Email: liping@math.pku.edu.cn

In capture-recapture studies, the most common method of constructing confidence interval for population size is by using the asymptotic normal property of the population size estimation. In this work we derive the confidence interval via the likelihood ratio test for two capture-recapture models. The large sample property is proved, and simulations are exploited to show that the proposed confidence interval is better than existing one. A real example is also analyzed.

Nested sub-sample search algorithm for estimation of threshold

Dong Li

Mathematical Sciences Center & Tsinghua Center for Statistics Science, Tsinghua University, Beijing, China
Email: dongli@math.tsinghua.edu.cn

Threshold models have been popular for modeling nonlinear phenomena in diverse areas, in part due to their simple fitting and often clear model interpretation. A commonly used approach to fit a threshold model is the (conditional) least squares method, for which the standard grid search typically requires O(n) operations for a sample of size n, which is substantial for large n, especially in the context of panel time series. This paper proposes a novel method, the *nested sub-sample search algorithm*, which reduces the number of least squares operations drastically to O(log n) for large sample size. We demonstrate its speed and reliability via Monte Carlo simulation studies with finite samples. Possible extension to maximum likelihood estimation is indicated.

Covariate-adaptive randomization with variable selection in high dimensional data

Jianxin Yin

Center for Applied Statistics and School of Statistics, Renmin University, Beijing, China
Email: jyin@ruc.edu.cn

In clinical trials, balancing treatment allocation for influential covariates is critical. On the other hand, the number of measured covariates is usually much larger than the number of recruited patients and growing very fast as the sample size increases, among which only a small fraction of them are really relevant to the given response. Recently, Hu and Hu (2012) proposed a new covariate-adaptive randomization procedure which can control three types of imbalance. However, they assume all the relevant covariates are known so they need not to select the important variables. How to select the potential important covariates among a diverging number of candidates and balance treatment allocation upon the selected variable set is the main subject of this paper. Under group sparsity assumption, we tackle this problem under the framework of multi-task learning via group-LASSO algorithm. From the variable selection and multi-task learning aspect, we relax the condition of equal sample sizes and allow unequaled sample sizes for different treatments. Compared to Lounici et al. (2011), we get a new result about the probability in the non-asymptotic oracle inequality which depend on the sample size. These two new results for multi-task learning are of independent interest. Finally we show under certain regulatory conditions, the regularized adaptive design method can control the asymptotic variance and select the true influential covariates set simultaneously. Simulation study has shown support to our theoretical discovery for the proposed method.

Keynote Session 4

- ***Chair: Xuegong Zhang***

Whole genome sequencing of six dog breeds from continuous altitudes reveals adaption to high-altitude hypox

Yixue Li

Shanghai Institute for Biological Science, Chinese Academy of Sciences, Shanghai, China
Email: yxli@sibs.ac.cn

The hypoxic environment imposes severe selective pressure on species living at high altitude. To understand the genetic bases of adaptation to high altitude in dogs, we performed whole-genome sequencing of 60 dogs including five breeds living at continuous altitudes along the Tibetan Plateau from 800 to 5100 m as well as one European breed. More than1503 sequencing coverage for each breed provides us with a comprehensive assessment of the genetic polymorphisms of the dogs, including Tibetan Mastiffs. Comparison of the breeds from different altitudes reveals strong signals of population differentiation at the locus of hypoxia-related genes including endothelial Per-Arnt-Sim (PAS) domain protein 1 (EPAS1) and beta hemoglobin cluster. Notably, four novel nonsynonymous mutations specific to high-altitude dogs are identified at EPAS1, one of which occurred at a quite conserved site in the PAS domain. The association testing between EPAS1 genotypes and blood-related phenotypes on additional high-altitude dogs reveals that the homozygous mutation is associated with decreased blood flow resistance, which may help to improve hemorheologic fitness. Interestingly, EPAS1 was also identified as a selective target in Tibetan highlanders, though no amino acid changes were found. Thus, our results not only indicate parallel evolution of humans and dogs in adaptation to high-altitude hypoxia, but also provide a new opportunity to study the role of EPAS1 in the adaptive processes.

Session 2: Life Science II

- ***Chair: Tingting Zhang***

Li Zhang

Peking Union Medical College Hospital, Beijing, China
Email: zhanglipumch@aliyun.com

TBD.

Big Data in China Healthcare System

Simeng Han

Analysis Group Inc., Beijing, China
Email: Simeng.Han@analysisgroup.com

There has been an increasing demand of scientific evidence based on real-world data in healthcare system. “Big data”, such as electronic medical records (EMRs), insurance claims data, registry data and pharmacy data have been playing important roles. These data contain rich demographics and clinical information, which can provide valuable information for potentially improving care of chronic diseases, uncovering the clinical effectiveness of treatments and reducing readmissions. Among all the data sources, 100% EMR data contains most accurate and comprehensive clinical and cost information including medications, procedures, lab values, and radiographic findings. In this talk, we will first present advantages and limitations of different data sources in China healthcare system. And then we will present a recent study in venous thromboembolism treatment pattern using EMR data to demonstrate its data elements, potential applications and challenges.

Exploring genetic and epigenetic data in cancer research: two case studies

Xiaoqi Zheng

Department of Mathematics, Shanghai Normal University, Shanghai, China
Email: zheng.shnu@gmail.com

1) Tumor impurity has been a major technical issue in tumor profiling studies. We propose a statistical algorithm MethylPurify that uses regions with bisulfite reads showing discordant methylation levels to infer tumor purity from tumor samples alone. With the purity estimates, MethylPurify can identify differentially methylated regions (DMRs) from individual tumor samples. It is the first computational method to estimate tumor purity and make differential DNA methylation calls from tumor methylome data alone, without genomic variation information or prior knowledge from other datasets. In simulations with mixed bisulfite reads from cancer and normal cell lines, MethylPurify correctly inferred tumor purity and identified over 96% of the

DMRs. On real patient data where tumor to normal comparison were used as golden standard, MethylPurify gave satisfactory DMR calls from tumor samples alone. Comparison with TCGA methylation results further suggests that DMRs called from tumor samples alone are equally accurate as the tumor to normal comparison, and included DMRs missed by the latter due to tumor heterogeneity.

2) Predicting the response of patients to a given therapy is a major goal in modern oncology that should ultimately lead to the personalized treatment. Existing methods on drug sensitivity prediction mainly use different kinds of genomic information as input features to regress or classify the response for a specific drug, while associations between different drugs and relationships between samples are neglected. In this work, we proposed a two-layer integrated network model (CSN: cell line similarity network and DAN: drug similarity network) for prediction of drug response against a cell line using a local weighted model. Using CCLE and CGP as benchmark datasets, our model achieved comparable prediction performance to the ongoing Elastic net model based only on one single layer of the integrated network. By using the whole integrated network, our final accuracy (Pearson correlation coefficient between predicted and observed drug responses) achieved around 0.6 for most drugs, which is significantly higher than that by the Elastic net model. As an application of the integrated network, we completed all missing drug response values in CGP dataset. Our model reasonably assigns much lower IC50 values for BRAF mutant cell lines than BRAF wild type cell lines for all three MEK inhibitors, which is consistent with known data and supported from the literatures.

Data mining and its Application in Medicine

Zhongyu Liu

The Anal-Colorectal Surgery Institute, No. 150 Central Hospital of Chinese PLA, Luoyang, China
Email: zhongyujohn@gmail.com

In my presentation, the usages and issues of data mining in medicine will be firstly introduced in brief; secondly, the characteristics of medical data and challenges of data mining in medicine will be stated; finally, we shall present the pre-process of medical data, the common data mining modeling, evaluations and application examples.

Session 6: Computer Science & Industry

❖ *Chair: Peter Qian*

Tensor dimension reduction method for chemical sensing

Wenxuan Zhong

University of Georgia
Email: wenxuan@uga.edu

With the rapid development of science and technology in the past decades, large amounts of tensor data are routinely collected, processed and stored and aected commercial activities nowadays. To address the statistical challenges that arise in analyzing tensor data, we proposed the tensor dimension reduction regression (TDR) model, which assumes a nonlinear dependency between a response variable and a projection of some tensor predictors. A novel sequential estimation approach, called SIDRA, has been proposed to estimate the projection. In this talk, I will briefly discuss the theoretical underpinning of SIDRA and demonstrate it application in chemical sensing. Preliminary studies demonstrate the great potential of SIDRA in improving the prediction accuracy in chemical sensing.

Big Data in Telco-Infrastructure, use case, and challenges

Lin Zhang

[1]School of Information and Communciation Engineering, Beijing University of Posts and Telecommunctions,
Xitucheng Raod No. 10, Haidian District, Beijing, China
Email: zhanglin@bupt.edu.cn

With the proliferation of powerful mobile devices and innovative networked applications, Internet are emerging to be the primary means for people to consume and share information. In order to support the explosive growth of data volume, players of Telco industry including network operators and service providers need to manage and plan their network and computing resources accordingly. Hence, the ability to accurately and extensively monitor and analyze the traffic data is a fundamental necessity for network operations and market optimization. Towards this end, Telco industry is starting to leverage big data technolgies to perform deep data anlysis. In this presentation, we will introduce: 1) A data infrastructure for network traffic analysis, which includes traffic record collector, file and data storage, analysis applications, accessing interface, and cluster management. 2) A use case to show how big data technologis can benifit the Telco industry, which is user profiling with explaination of methodologies and analysis resutls, 3) Chellenges of applying big technolgies in Telco we found in our works, including the impact of the huge volume of traffic data produced in extra high speed network link, the requirment of an unified infrastructure for batch, streaming and iterative data processing, and the critical requirement of ad-hoc query on big data with near real-time criteria.

Bayesian Aggregation of Classifiers with Various Qualities

Weifeng Zhang

Nanjing University of Posts and Telecommunctions, Nanjing, China
Email: wfbreezee@gmail.com

In many practical classification problems, we often have results from multiple competing classifiers for the same data set. In this work, we present a Bayesian approach, called Bayesian Classifier Aggregation (BCA), which overcomes the difficulties how to distinguish high quality base classifiers from low quality ones and treat them differently. By attaching two quality parameters to each base classifier, one for sensitivity and one for specificity, and estimating these parameters along with the aggregation process, BCA can measure the quality of base classifiers in a quantitative way, and improve the aggregation with the information. Both simulation experiments and real data applications show that BACD outperforms existing methods.

Computer experimental design for metamaterial

Chunlin Ji

Kuang-Chi Institue of Advanced Technology, Shenzhen, China
Email: chunlin.ji@kuang-chi.org

We discuss some computer experimental design issues in metamaterial. Metamaterial is a new emerging area in physics and electronic community. Metamaterial has numerous numbers of elements (also called unit cell), which may have different electromagnetic (EM) response. With careful design of the unit cell and the design of the spatial distribution of these cells, the metamaterial can have fantastic EM response. In this talk, we share our experience in building the metamodel for the complex EM response curve of each unit. Particularly, we proposed some novel idea to deal with complex response, i.e. curves, in the design of computer experiment. We demonstrate our proposed method by some non-trivial examples.

Keynote Session 5

- ***Chair: Songxi Chen***

Regime-Switching Factor Models for High-Dimensional Time Series

Rong Chen

Department of Statistics, Rutgers University, Piscataway, NJ 08854, USA
Email: rongchen@stat.rutgers.edu

We consider a factor model for high-dimensional time series with a regime switching dynamics.The switching is assumed to be driven by an unobserved Markov chain; the mean, factor loading matrix and covariance matrix of the error process are different among the regimes. The model is an extension of the traditional factor models for time series and provides flexibility in dealing with real applications in which underlying states may change over time. We propose an iterative approach to estimate the loading spaces of each regime as well as the states of the hidden Markov chain at each time, by combining eigenanalysis and Viterbi algorithm. The theoretical properties of the procedure are investigated. Simulation results and analysis of a real example are presented.

Session 7: Statistical Theory and Methodology III

- ***Chair: Qunhua Li***

Hewei Pang

China Academy of Space Technology
Email: zhanglipumch@aliyun.com

TBD.

New series of sliced space-filling designs for computer experiments

Mingyao Ai

School of Mathematical Sciences, Peking University, Beijing, China
Email: myai@pku.edu.cn

Latin hypercube designs have been widely used in computer experiments with quantitative factors. For conducting multiple computer experiments with different levels of accuracy, nested space-filling designs have been proposed to deal with them. When there are both qualitative and quantitative factors in computer experiments, sliced space-filling designs have been proposed to conduct such experiments. Here we present a general framework for constructing sliced space-filling designs for more flexible parameters of designs in which the whole design and each slice not only achieve maximum stratification in univariate margins, but also achieve stratification in two- or more-dimensional margins. Compared with other designs, the new constructed designs have better space-filling property or have more columns. The construction is based on a new class of sliced orthogonal arrays, called balanced sliced orthogonal arrays. Several approaches to constructing such balanced sliced orthogonal arrays under different level-collapsing projections are developed.

A Generalized OEM Algorithm for Large-Scale Statistical Computation

Peter Qian

University of Wisconsin-Madison, Madison, USA
Email: peterq@stat.wisc.edu

We propose an iterative algorithm that is highly efficient in large-scale statistical computation settings. Our generalized OEM (GOLEM) algorithm, a modification of Orthogonalizing EM (OEM), enjoys fast convergence for ordinary least squares (OLS) problems and is also easily extended to regularized least squares problems including penalties such as the lasso, ridge, and SCAD. In our reformulation, we provide a different geometric interpretation of the OEM algorithm and a computationally efficient solver based on the new geometric perspective. An optimal convergence rate is also proved based on this perspective. In addition, we extend the OEM technique to more general settings such generalized linear models, resulting in an algorithm that is state-of-the-art for logistic regression and regularized logistic regression in settings where $n >> p$. In general, our alternate formulation of OEM for OLS has computational complexity $O(np)$. We prove that GOLEM converges for convex losses and converges to a local solution when the loss function is nonconvex. We demonstrate the state-of-the-art performance of GOLEM on several large-scale datasets, both real and synthetic.

A Multilevel Radial-Basis-Function Method for Computer Experiments

Rui Tuo

Chinese Academy of Sciences, Beijing, China
Email: tuorui@amss.ac.cn

The modeling and analysis for multi-fidelity computer experiments have received increasing attention recently. It has been shown that better prediction can be achieved by using multi-fidelity data than single-fidelity data. However, mathematical theory on the statistical analysis for multi-fidelity computer models has not been established yet. In this work we propose a multilevel collocation method for multi-fidelity computer experiments using the radial basis functions as the interpolator. Theoretical analysis shows that the proposed method possesses a much higher convergence rate than the single-level method. The surrogate model of the proposed method is similar with the auto-regressive model suggested by Kennedy and O'Hagan (2000). This connection can give a mathematical foundation for the Kenendy-O'Hagan's method. Moreover, our analysis provides optimal experimental designs for such models. The proposed method is applicable for uncertainty quantification, prediction and optimization problems.

Keynote Session 6

- ***Chair: Hansheng Wang***

Leveraging in Big Data Regression

Ping Ma

Department of Statistics, University of Georgia, USA
Email: pingma@uga.edu

Advances in science and technology in the past a few decades have led to big data challenges across a variety of fields. Extraction of useful information and knowledge from big data has become a daunting challenge to both the science community and entire society. To tackle this challenge requires major breakthroughs in efficient computational and statistical approaches to big data analytics.

In this talk, I will present some leveraging algorithms, which make a key contribution to resolving the grand challenge. In these algorithms, by sampling a very small representative sub-dataset using smart algorithms, one can effectively extract relevant information of vast data sets from the small sub-dataset. Such algorithms are scalable to big data. These efforts allow pervasive access to big data analytics especially for those who cannot directly use supercomputers. More importantly, these algorithms enable massive ordinary users to analyze big data using tablet computers.

Session 8: Statistical Theory & Methodology IV

- ***Chair: Junni Zhang***

Bayesian Variable Selection for Linear Regression with Interaction Terms

Wensheng Zhu

Northeast Normal University, Changchun, Chia
Email: wszhu@nenu.edu.cn

In modern statistics, we always encounter high-throughput data with huge number of covariates or features for a small number of subjects. That's what we say "p > n". In many cases, researchers show a great interest in the interactions of these covariates. However, it is a challenging task to construct the suitable model by selecting the active covariates among tens of thousands of interactions. In this talk, we propose a Bayesian variable selection approach to identify interactions in the presence of huge dimensional covariates. Spike and slab Gaussian priors are used on the main effects as well as interactions, which shrink and diffuse, respectively as the ample size increases. We show the strong selection consistency of the proposed method in the sense that the posterior probability of the true model converges to 1 for p=exp(o(n)). In practice, we choose the model with the highest posterior probability, which can be achieved through posterior sampling with a Gibbs sampler. Simulations in genetic association studies indicate that our proposed method offers merits in the detection of gene-gene and gene-environmental interactions.

Integrative and Regularized Principal Component Analysis of Multiple Sources of Data

Binghui Liu

Northeast Normal University, Changchun, China
Email: wszhu@nenu.edu.cn

Integration of data of disparate types has become increasingly important to enhancing the power for new discoveries by combining complementary strengths of multiple types of data. One application is to uncover tumor subtypes in human cancer research, in which multiple types of genomic data are integrated, including gene expression, DNA copy number and DNA methylation data. In spite of their successes, existing approaches based on joint latent variable models require stringent distributional assumptions and may suffer from unbalanced scales (or units) of different types of data and non-scalability of the corresponding algorithms. In this paper, we propose an alternative based on integrative and regularized principal component analysis, which is distribution-free, computationally efficient, and robust against unbalanced scales. The new method performs dimension reduction simultaneously on multiple types of data, seeking

data-adaptive sparsity and scaling. As a result, in addition to feature selection for each type of data, integrative clustering is achieved. Numerically, the proposed method compares favorably against its competitors in terms of accuracy (in identifying hidden clusters), computational efficiency, and robustness against unbalanced scales. In particular, compared to a popular method, the new method was competitive in identifying tumor subtypes associated with distinct patient survival patterns when applied to a combined analysis of DNA copy number, mRNA expression and DNA methylation data in a glioblastoma multiforme study.

Statistical Learning from Large Molecular Dynamics Datasets

Xiaodan Fan

The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China
Email: xfan@sta.cuhk.edu.hk

Molecular Dynamics (MD) data analysis is one of the earliest big data problems. The data is distributedly generated by thousands of computers, each of which ran an MD simulation with a unique initial state. The task of statistical learning on these datasets is to elucidate the energy landscape of a protein by integrating thousands of MD trajectories. We will use both genometric distance and dynamic distance to probe the grouping structure of conformations within the landscape.

A Dynamical System Approach to Variable Selection

Yuan Yao

Peking University, Beijing, China
Email: yaoy@math.pku.edu.cn

We approach variable selection in high dimensional statistics or sparse signal recovery from their noisy linear measurements by solving nonlinear differential inclusions. Such dynamics admits an extremely simple implementation which exhibits statistical properties only achievable by non convex regularization. Precisely we show that under proper conditions, there exists a bias-free and sign-consistent point on their solution paths, which corresponds to a signal that is the unbiased estimate of the true signal and whose entries have the same signs as those of the true signs. Therefore, their solution paths are regularization paths better than the LASSO regularization path, since the points on the latter path are biased. Theoretical guarantees such as sign-consistency and minimax optimal l_2-error bounds are established in both continuous and discrete settings for specific points on the paths. Early-stopping regularization is necessary for identifying these points. This is a joint work with Stan Osher, Feng Ruan, Jiechao Xiong and Wotao Yin.

Session 9: Statistics & Social Science

❖ *Chair: Xiaodan Fan*

Natural language Processing-related Research at Tsinghua University

Maosong Sun

Department of Computer Science, Tsinghua University, Beijing, China
Email: sms@tsinghua.edu.cn

This talk briefly introduces NLP-related research, with emphasis on statistical methodologies, at Tsinghua University, including Chinese word segmentation and part-of-speech tagging, keyword extraction for microblogs, word sense disambiguation, as well as machine translation. Some of suggestions for the future work are proposed, for instance, NLP in terms of naturally annotated big data.

Bio: Maosong Sun is a professor of the Department of Computer Science and Technology of Tsinghua University. His research interests are natural language processing, Web intelligence and social computing. He has published over 150 papers in academic journals and international conferences (including Computational Linguistics, IEEE Intelligent Systems, ACM TALIP, Journal of Quantitative Linguistics, IJCAI, AAAI, ACL, EMNLP, COLING, and VLDB). He has served as program committee members in numerous national and international conferences, and many times as conference chairs or program committee chairs. Prof. Sun is the vice president of Chinese Information Processing Society of China, the council member of China Computer Federation, the council member of Chinese Association for Artificial Intelligence, the Editor-in-chief of the Journal of Chinese Information Processing, the director of Tsinghua University-National University of Singapore Joint Research Centre on Extreme Search Technology, as well as the director of Large-scale Online Education Centre of Tsinghua University.

Statistics and Digital Humanities

Jing Chen

Art Institute, Nanjing University & School of Humanities, Rice University
Email: ccjj2008@gmail.com

Cross-disciplinary researchers and institutions around the world are pioneering ways to merge information technology and humanities' studies. The Digital Humanities (DH), formerly "Humanities Computing," actually signals a profound change in humanities' research methods, in the questions humanists are able to pursue. Some of these changes are obvious. Providing algorithmic approaches to large humanities data sets and subjecting them to content/data analysis and visualization; incorporating geospatial data into classroom projects; building new ways to see things, manipulate images, distribute them and communicate with learners: these changes open pedagogic possibilities that lie beyond the capacity of the individual working in a library or

archive.

Among the methodologyies applied for the peojects of digital humanities, statistical analysis is almost the most popolur one and has been generally used for the content analysis and the visualization of the data collection. Information collected through historical records, perusal of archival documents, semi-structured interviews, unstructured interviews, study of photographs and maps, and other methods that might initially yield non-numerical information can often be coded into categories that can subsequently be statistically manipulated. However, difficulty lies that the meaning of what has been observed derives not from "counting" something but rather from understanding how to interpret what was observed, with the counts helping to judge the strength of the interpretation. The interpretation depends upon a researcher's understanding of the social context for what was observed. Meanwhile, some texts or documents of humanities studies can not be directly coded, classified, and categorized and then "counted" in a quatitative way. So how to think about the role of stastistics in the digital humanities projects? How to expand the application possiblities of statistics in the digital humanties studies?

My presentation will start from the breif introduction to the history and definition of digital humanties and then talk about the most popular applications of statistics in digital humanities projects with some sucessful cases. I will explain the possibilities of using statistical analysis in my own research on Chinese Commercial Advertisements in Modern China period. Also I hope I will get feedback and comment from other scholars on the questions I am facing now during applying the stastitical analysis to the historical studies of images of advertisements and the contextual information related to the moment and the place, when and where advertisements emerged.

Statistical Approaches for Text Mining in Big Data Era

Ke Deng

Mathematical Sciences Center, Tsinghua University, Beijing, China
Email: kdeng@math.tsinghua.edu.cn

Discovering patterns and knowledge from a set of text is an important problem in many disciplines such as biomedical research, linguistics, artificial intelligence and sociology. However, texts from different research domains usually have very different properties, and it is often unrealistic to base text analysis on supervised approaches where training data are needed. We propose here a novel unsupervised approach for mining Chinese texts, which can achieve word/phrase discovery and text segment simultaneously. Combine the method with other domain knowledge show great power to solve various practical problems.

Xinshu Zhao

Hong Kong Baptist University, Hong Kong
Email: zhao@hkbu.edu.hk

TBD.