THE HONG KONG POLYTECHNIC UNIVERSITY 香港理工大學

Academy of Mathematics and Systems Science

DEPARTMENT OF APPLIED MATHEMATICS 應用數學系

The Workshop on Youth Statistician Forum: Prospect and Perspective in Statistical Inference and its Application

25-26 June 2015

Objective:

This workshop aims to call local and overseas youth statisticians to share their insight into current and future statistical development and challenge particularly in big data field. It provides a channel to communicate and promote future academic interactivities among overseas and local youth statisticians in a wide range of statistically related fields.

Invited speakers

Chuanhai Liu	Purdue University, USA
Xiaolin Chen	China University of Petroleum, China
Bingqing Lin	Shenzhen University, China
Samuel Muller	The University of Sydney, Australia
Qiang Wu	Middle Tennessee State University, USA
Hongqi Xue	University of Rochester, USA
Zhisheng Ye	National University of Singapore, Singapore
Yiming Ying	University at Albany, State University of New York, USA
Tao Zhang	Guangxi University of Science and Technology, China
Xiaodan Fan	The Chinese University of Hong Kong
Xin Guo	The Hong Kong Polytechnic University
Catherine Liu	The Hong Kong Polytechnic University
Zhen Pang	The Hong Kong Polytechnic University
Tony Sit	The Chinese University of Hong Kong
Tiejun Tong	Hong Kong Baptist University
Junhui Wang	City University of Hong Kong
Yingying Wei	The Chinese University of Hong Kong
Can Yang	Hong Kong Baptist University
Chun Yip Yau	The Chinese University of Hong Kong
Aijun Zhang	Hong Kong Baptist University
Phillip Yam/Zheng Zhang	The Chinese University of Hong Kong

Date/Venue: 25 June 2015 (Thu)/ Y302, Lee Shau Kee Building, PolyU
26 June 2015 (Fri)/M1603, Li Ka Shing Tower, PolyU

Enquiry: Miss Eunice Hung; Tel: (852) 3400 3908; Email: eunice.hung@polyu.edu.hk

Sponsors: The AMSS-PolyU JRI and The Hong Kong Polytechnic University

Website: http://www.polyu.edu.hk/ama/amss_jri

All are welcome

The Workshop on Youth Statistician Forum: Prospect and Perspective in Statistical Inference and its Application

25-26 June 2015, The Hong Kong Polytechnic University

Program and Abstracts

Date and Venue:
25 June 2015 (Thu) / Y302, Lee Shau Kee Building, PolyU
26 June 2015 (Fri) / M1603, Li Ka Shing Tower, PolyU

Sponsors:
The AMSS-PolyU JRI and The Hong Kong Polytechnic University

Organizers:
Drs. Xin Guo
Catherine Liu
Zhen Pang
Department of Applied Mathematics

Objective

This workshop aims to call overseas and local youth statisticians to share their insight into current and future statistical development and challenge particularly in big data field. It provides a channel to communicate and promote future academic interactivities among overseas and local youth statisticians in a wide range of statistically related fields.

Invited Speakers

Overseas

Chuanhai Liu	Purdue University, USA
Xiaolin Chen	Chinese University of Petroleum, China
Bingqing Lin	Shenzhen University, China
Samuel Müller	The University of Sydney, Australia
Qiang Wu	Middle Tennessee State University, USA
Hongqi Xue	University of Rochester, USA
Zhisheng Ye	National University of Singapore, Singapore
Yiming Ying	University at Albany, State University of New York, USA
Tao Zhang	Guangxi University of Science and Technology, China

Local

Xiaodan Fan	The Chinese University of Hong Kong
Xin Guo	The Hong Kong Polytechnic University
Catherine Liu	The Hong Kong Polytechnic University
Zhen Pang	The Hong Kong Polytechnic University
Tony Sit	The Chinese University of Hong Kong
Tiejun Tong	Hong Kong Baptist University
Junhui Wang	City University of Hong Kong
Yingying Wei	The Chinese University of Hong Kong
Can Yang	Hong Kong Baptist University
Chun Yip Yau	The Chinese University of Hong Kong
Aijun Zhang	Hong Kong Baptist University
Phillip Yam/Zheng Zhang	The Chinese University of Hong Kong

Program

FIRST DAY [Venue: Y302, see campus map attached]

9:00-9:25
REGISTRITION

9:25-9:45
OPENING: Welcome words from the Department Head, Prof. Xiaojun Chen
GROUP PHOTO

SESSION I **Chair: Catherine Liu**

9:45-10:30
PLENARY TALK
Chuanhai Liu, Purdue University [refer to page 8]
Inferential Models: A New School of Thought on Scientific Inference for Next Generations

10:30-10:45
TEA BREAK

SESSION II **Chair: Yiming Ying**

10:45-11:10
Hongqi Xue, University of Rochester [refer to page 12]
Numerical Error and Measurement Error in Statistical Analysis for Ordinary Differential Equation Models

11:10-11:35
Xiaodan Fan, The Chinese University of Hong Kong [refer to page 6]
Some Thoughts on Probabilistic Data Integration

11:35-12:00
Samuel Müller, The University of Sydney [refer to page 9]
Model Selection with Mplot

12:00-13:15
LUNCH (Staff Club, 5/F, Communal Building, see attached map)

SESSION III **Chair: Samuel Müller**

13:15-13:40
Tony Sit, The Chinese University of Hong Kong [refer to page 10]
Accelerated Failure Time Model under General Biased Sampling Scheme

13:40-14:05
Zhisheng Ye, National University of Singapore [refer to page 14]
Augmenting the Unreturned for Field Data with Information on Returned Failures Only

14:05-14:30
Tao Zhang, Guangxi University of Science and Technology [refer to page 16]
An Extended Single Index Model with Missing Response at Random

14:30-14:45
TEA BREAK

SESSION IV **Chair: Tiejun Tong**

14:45-15:10
Junhui Wang, City University of Hong Kong [refer to page 11]
Classification with Unstructured Predictors with an Application to Sentiment Analysis

15:10-15:35
Chun Yip Yau, The Chinese University of Hong Kong [refer to page 14]
High Order Bias Corrected Estimator for Time-Average Variance Constant

15:35-16:00
Xin Guo, The Hong Kong Polytechnic University [refer to page 7]
The Local Edge Machine: Inference of Dynamic Models of Gene Regulation

16:00-16:25
Yiming Ying, University at Albany, State University of New York [refer to page 15]
Online Pairwise Learning Algorithms (OPERA)

16:25-16:40
TEA BREAK

SESSION V **Chair: Xiaodan Fan**

16:40-17:05
Tiejun Tong, Hong Kong Baptist University [refer to page 10]
Shrinkage-Based Diagonal Hotelling's Tests for High-Dimensional Small Sample Size Data

17:05-17:30
Phillip Yam/Zheng Zhang, The Chinese University of Hong Kong [refer to page 16]
Globally Efficient Nonparametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting

17:30-17:55
Catherine Liu, The Hong Kong Polytechnic University [refer to page 8]
Testing Equality of Covariance Operators/Matrices For Functional/High-Dimensional Data

18:15, DINNER
港潮樓，尖沙咀加連威老道 100 號港晶中心 1 樓(see map attached)

THE END OF THE FIRST DAY

SECOND DAY [Venue: M1603, see map attached]

SESSION VI **Chair: Zhen Pang**

9:30-9:55
Qiang Wu, Middle Tennessee State University [refer to page 12]
Consistency Analysis of the Minimum Error Entropy Algorithm

9:55-10:20
Yingying Wei, The Chinese University of Hong Kong [refer to page 11]
A Scalable Integrative Model for Heterogeneous Genomic Data Types under Multiple Conditions

10:20-10:45
Can Yang, Hong Kong Baptist University [refer to page 13]
IMAC: A Flexible Statistical Approach to Integrating Multilayered Annotation for Characterizing Functional Roles of Genetic Variants that Underlie Human Complex Phenotypes

10:45-11:00
TEA BREAK

SESSION VII **Chair: Xin Guo**

11:00-11:25
Aijun Zhang, Hong Kong Baptist University [refer to page 15]
Big Data Analytics in Online Education

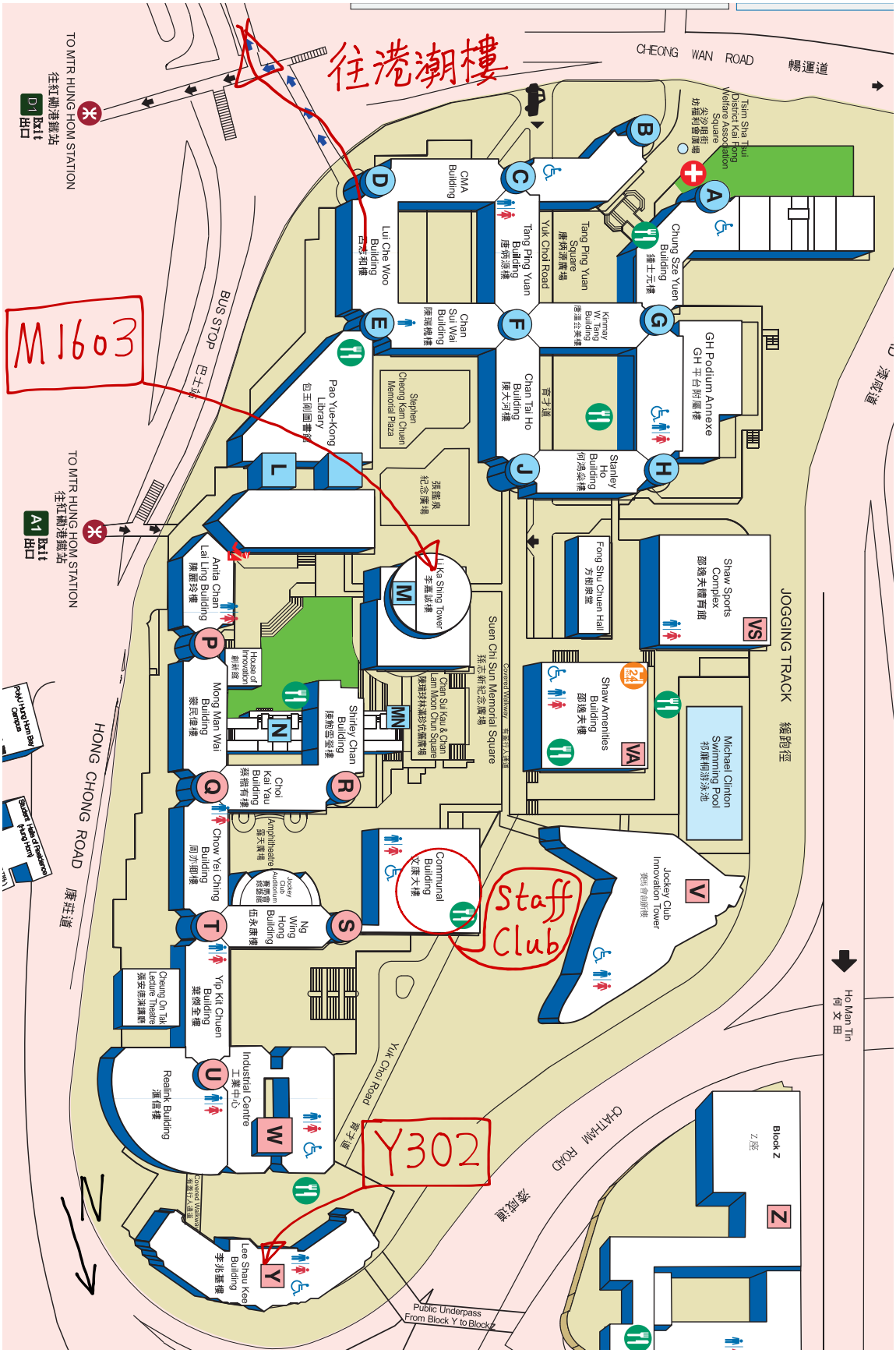
11:25-11:50
Bingqing Lin, Shenzhen University [refer to page 7]
LFCseq: a Nonparametric Approach for Differential Expression Analysis of RNA-Seq Data

11:50-12:15
Xiaolin Chen, China University of Petroleum [refer to page 6]
Quantile Correlation Screening for Ultrahigh Dimensional Heterogeneous Data

12:15-12:40
Zhen Pang, The Hong Kong Polytechnic University [refer to page 9]
Cluster Feature Selection in High Dimensional Linear Models

13:00, LUNCH
港潮樓，尖沙咀加連威老道 100 號港晶中心 1 樓(see map attached)
THE END OF THE SECOND DAY

Attached map 1, campus



Attached map 2, 港潮樓，尖沙咀加連威老道 100 號港晶中心 1 樓

Abstracts

Quantile Correlation Screening for Ultrahigh Dimensional Heterogeneous Data

Xiaolin Chen, China University of Petroleum

In this talk, we systematically investigate the quantile correlation sure independence screening (QC-SIS) for ultrahigh dimensional heterogeneous data. We unveil the question under which assumptions and correlation structures QC-SIS enjoys the ranking consistency and sure screening properties when the number of features grows at an exponential rate of the sample size. Due to the nature of marginal independence screening, QC-SIS may miss truly active features which are marginally independent of the response, but contribute to the response jointly with other predictors. In addition, when there are many irrelevant predictors which are highly correlated with some strongly active predictors, QC-SIS may fails to identify other active predictors with relatively weak marginal signals. To enhance the finite sample performance, an iterative quantile correlation sure independence screening (QC-ISIS) is suggested based on the partial quantile correlation. Extensive simulation studies are carried out to examine the performances of QC-SIS and QC-ISIS with its main competitors. We also illustrate the proposed QC-SIS and QC-ISIS by real data examples.

Joint work with Dr. Chunling Liu and Prof. Kam Chuen Yuen.

Some Thoughts on Probabilistic Data Integration

Xiaodan Fan, The Chinese University of Hong Kong

Many scientific truths can be probed from multiple available datasets produced by different research groups. How to weight their relative trustworthiness is a common challenge. This problem becomes more and more common in the ear of big data. Probabilistic models for integrating these datasets can be a principled way for weighting, but there are many unsolved aspects. Based on our experience on integrative clustering for large heterogeneous data sets, we will discuss on the prospects and challenges of probabilistic data integration.

The Local Edge Machine: Inference of Dynamic Models of Gene Regulation

Xin Guo, The Hong Kong Polytechnic University

A gene regulatory network is a collection of genes that regulate each other, through RNA and protein expression products. Gene regulatory networks enable organisms to predict and adapt to environment changes. Understanding the regulatory relationship is a big challenge in molecular biology and systems biology. We developed the Local Edge Machine (LEM), which is an algorithm to infer the network from temporally dynamic gene expression data. LEM uses differential equation systems with the Hill function model to fit the expression level data, and is regularized with preference on robust systems. In validation studies on both *in silico* and *in vivo* data, our method outperforms previously reported methods by wide margins.

LFCseq: A Nonparametric Approach for Differential Expression Analysis of RNA-Seq Data

Bingqing Lin, Shenzhen University

Background:

With the advances in high-throughput DNA sequencing technologies, RNA-seq has rapidly emerged as a powerful tool for the quantitative analysis of gene expression and transcript variant discovery. In comparative experiments, differential expression analysis is commonly performed on RNA-seq data to identify genes/features that are differentially expressed between biological conditions. Most existing statistical methods for differential expression analysis are parametric and assume either Poisson distribution or negative binomial distribution on gene read counts. However, violation of distributional assumptions or a poor estimation of parameters often leads to unreliable results.

Results:

In this paper, we introduce a new nonparametric approach called LFCseq that uses log fold changes as a differential expression test statistic. To test each gene for differential expression, LFCseq estimates a null probability distribution of count changes from a selected set of genes with similar expression strength. In contrast, the nonparametric NOISeq approach relies on a null distribution estimated from all genes within an experimental condition regardless of their expression levels.

Conclusion:

Through extensive simulation study and RNA-seq real data analysis, we demonstrate that the proposed approach could well rank the differentially expressed genes ahead of non-differentially expressed genes, thereby achieving a much improved overall performance for differential expression analysis.

Testing Equality of Covariance Operators/Matrices For Functional/High-Dimensional Data

Catherine Liu, The Hong Kong Polytechnic University

The purpose of this talk is twofold. The first goal is to propose a unified methodology for testing equality of covariance operators of two functional samples no matter if the functional data are dense or sparse, regular or irregularly spaced. To demonstrate this, a two-step procedure is presented which leads to a global testing statistic. The second goal is, from the insight of functional data analysis, to develop a novel method to test equality of two high-dimensional covariance matrices. It might be an inchoate trying to apply the idea of functional data analysis into high-dimensional data study.

Under null and alternative hypotheses, asymptotic distributions of the testing statistics have been derived for afore two types of data. Extensive simulation experiments have been conducted. This indicates that the proposed approaches outperform existing work in terms of either size or power for testing problems both for functional data and high-dimensional data. Air pollution data in western south district of China is analyzed to illustrate our procedure of testing equality of covariance operators for functional data samples; Mitochondrial calcium concentration data is analyzed to demonstrate how our proposed method can be applied to test equality of covariance matrices for high-dimensional data.

Joint work with Mr. Jin Yang and Dr. Tao Zhang.

Inferential Models: A New School of Thought on Scientific Inference for Next Generations

Chuanhai Liu, Purdue University

As the tool for the science that converts experience, in the form of observed data, to knowledge about unknown quantities of interest, statistics will be fundamental to the ultimate success of scientific research. Developing solid foundations for scientific inference is the most fundamental but unsolved problem in statistics. We argue for two new basic principles, namely, the validity and efficiency principles, for truly prior-free probabilistic inference. With a brief introduction to a principle-based framework, called Inferential Models (IMs), in this talk we focus on demonstrating how IMs can provide deep understanding of prior-probabilistic inference for combining information and parameters of interest, called Conditional IMs and Marginal IMs. We conclude the talk with a list of open problems.

Model Selection with Mplot

Samuel Müller, The University of Sydney

This talk focuses on the computational aspects of selection criteria that are based on either inclusion or exclusion frequencies. We have developed the mplot R package which provides a collection of functions to aid exploratory variable selection. The package contains fast routines to make available modified versions of the simplified adaptive fence procedure (Jiang et al., 2009, Annals of Statistics) as well as other graphical tools such as variable inclusion plots and model selection curves (Mueller and Welsh, 2010, International Statistical Review; Murray et al, 2013, Statistics in Medicine). A browser based graphical user interface is provided to facilitate interaction with the results. These variable selection methods rely heavily on bootstrap resampling techniques. Fast performance for standard linear models is achieved using the branch and bound algorithm provided by the leaps package. The graphical model selection methods in mplot visualise popular model selection criteria that involve minimizing a penalized function of the data over a typically very large set of models. The penalty in the criterion function is controlled by a tuning parameter which determines the properties of the procedure. The implemented methods in mplot allow us to better explore the stability of model selection criteria through model selection curves and this is demonstrated through case studies.

Joint work with Garth Tarr and AH Welsh.

Cluster Feature Selection in High Dimensional Linear Models

Zhen Pang, The Hong Kong Polytechnic University

This talk concerns with variable screening when highly correlated variables exist in high dimensional linear models. The elastic net procedure (Zou and Hastie, 2005) which was designed for this situation may select the highly correlated variables, but include too many truly irrelevant variables. We propose a novel cluster feature selection (CFS) procedure based on the elastic net and linear correlation variable screening to enjoy the benefits of the two methods. When calculating the correlation between the predictor and the response, we consider the highly correlated group of the predictors instead of the individual ones. This is in contrast to the usual linear correlation variable screening. Within each correlated group, we apply the elastic net to select and estimate the variables. This avoids the drawback of mistakenly eliminating true non-zero coefficients for highly correlated variables like LASSO (Tibshirani, 1996) does. After applying the cluster feature selection procedure, maximum absolute sample correlation coefficient between clusters becomes smaller and any common model selection methods like SIS (Fan and Lv, 2008) or LASSO can be applied to improve the results. Extensive numerical examples including pure simulation examples and semi-real examples are conducted to show the good performances of our procedure.

Accelerated Failure Time Model under General Biased Sampling Scheme

Tony Sit, The Chinese University of Hong Kong

Right-censored time-to-event data are sometimes observed from a (sub) cohort of patients whose survival times can be subject to outcome-dependent sampling schemes. In this paper, we propose a unified estimation method for semiparametric accelerated failure time models under general biased estimating schemes. The proposed estimator of the regression covariates is developed upon a bias-offsetting weighting scheme and is proved to be consistent and asymptotically normally distributed. Large sample properties for the estimator are also derived. Using rank-based monotone estimating functions for the regression parameters, we find that the estimating equations can be easily solved via convex optimisation. The methods are confirmed through simulations and illustrated by application to real data sets on various sampling scheme including length-bias sampling, the case-cohort design and its variants.

Joint work with Jane Paik Kim and Zhiliang Ying.

Shrinkage-Based Diagonal Hotelling's Tests for High-Dimensional Small Sample Size Data

Tiejun Tong, Hong Kong Baptist University

High-dimensional small sample size data such as microarrays bring novel tools and also statistical challenges to genetic research. In addition to detecting differentially expressed genes, testing the significance of gene sets or pathway analysis has been recognized as an equally important problem. Owing to the "large p small n" paradigm, the traditional Hotelling's T2 test suffers from the singularity problem and therefore is not valid in this setting. In this paper, we propose a shrinkage-based diagonal Hotelling's test for both one-sample and two-sample cases. We also suggest several different ways to derive the approximate null distribution under different scenarios of p and n for our proposed shrinkage-based test. Simulation studies show that the proposed method performs comparably to existing competitors when n is moderate or large, but it is better when n is small. In addition, we analyze four gene expression data sets and they demonstrate the advantage of our proposed shrinkage-based diagonal Hotelling's test.

Classification with Unstructured Predictors with an Application to Sentiment Analysis

Junhui Wang, City University of Hong Kong

Unstructured data refers to information that lacks certain structures and cannot be organized in a predefined fashion. Unstructured data involve heavily on words, texts, graphs, objects or multimedia types of files that are difficult to process and analyze by traditional computational tools and statistical methods. In this talk, I will discuss ordinal classification with unstructured predictors and ordered class categories, where imprecise information concerning strengths between predictors is available for predicting the class labels. We integrate the imprecise predictor relations into linear relational constraints over classification function coefficients, where large margin ordinal classifiers are introduced, subject to quadratically many linear constraints. The proposed methods are implemented via a scalable quadratic programming algorithm based on sparse word representations. The advantage is demonstrated in a variety of simulated experiments as well as one large-scale sentiment analysis example on TripAdvisor.com customer reviews. If time permits, the asymptotic properties will also be discussed, which confirm that utilizing relationships among unstructured predictors can significantly improve prediction accuracy.

A Scalable Integrative Model for Heterogeneous Genomic Data Types under Multiple Conditions

Yingying Wei, The Chinese University of Hong Kong

A key problem in biology is how the same copy of a genome within a person can give rise to hundreds of cell types. Plentiful convincing evidence indicates multiple elements, such as transcription factor binding, histone modification, and DNA methylation, all contribute to the regulation of gene expression levels in different cell types. Therefore, it is crucial to understand how these heterogeneous regulatory elements collaborate together, how the cooperation at a given genomic region changes across diverse cell lines, as well as how such dynamic cooperation patterns across cell lines vary along the whole genome. Here, we propose a scalable hierarchical probabilistic generative model to cluster genomic regions according to the dynamic changes of their open chromatin and DNA methylation status across cell types. The model will overcome the exponential growth of parameter space as the number of cell types integrated increases. The fitted results of the model will provide a genome-wide region-specific, cell-line-specific open chromatin and DNA methylation landscape map.

Joint work with Mai Shi.

Consistency Analysis of the Minimum Error Entropy Algorithm

Qiang Wu, Middle Tennessee State University

Information theoretical learning (ITL) is an important research area in signal processing and machine learning. It uses concepts of entropies and divergences from information theory to substitute the conventional statistical descriptors of variances and covariances. The empirical minimum error entropy (MEE) algorithm is a typical approach falling into this framework and has been successfully used in both regression and classification problems. In this talk, I will discuss the consistency analysis of the MEE algorithm. For this purpose, we introduce two types of consistency. The error entropy consistency requires the error entropy of the learned function to approximate the minimum error entropy. It holds when the bandwidth parameter tends to 0 at an appropriate rate. The regression consistency requires the learned function to approximate the regression function. We proved that the error entropy consistency implies the regression consistency for homoskedastic models where the noise is independent of the input variable. But for heteroskedastic models, a counterexample is constructed to show that the two types of consistency are not necessarily coincident. A surprising result is that the regression consistency holds when the bandwidth parameter is sufficiently large. The regression consistency of two classes of special models is shown to hold with fixed bandwidth parameter. These results illustrate the complication of the MEE algorithm.

Numerical Error and Measurement Error in Statistical Analysis for Ordinary Differential Equation Models

Hongqi Xue, University of Rochester

We consider parameter estimation for nonlinear ordinary differential equation (ODE) models where analytic closed-form solutions are not available. The numerical solution-based nonlinear least squares (NLS) estimator is proposed. A numerical algorithm such as the Runge-Kutta algorithm is used to approximate the ODE solution. The asymptotic properties are established for the proposed estimators with consideration of both numerical error and measurement error. Our results show that if the maximum step size of the numerical algorithm goes to zero faster than a special rate, which depends on the order of the ODEs, then the numerical error is negligible compared to the measurement error. This provides a theoretical guidance in selection of the step size for numerical evaluations of ODEs. We illustrate our approach with both simulation studies and clinical data on HIV viral dynamics. Finally, we extend the above model and method to their generalized ODE versions for fitting discrete data.

IMAC: a Flexible Statistical Approach to Integrating Multilayered Annotation for Characterizing Functional Roles of Genetic Variants that Underlie Human Complex Phenotypes

Can Yang, Hong Kong Baptist University

Recent international projects, such as the Encyclopedia of DNA Elements (ENCODE) project, the Roadmap project and the Genotype-Tissue Expression (GTEx) project, have generated vast amounts of genomic annotation data measured at the multiple layers, e.g., epigenome and transcriptome. These multilayered annotation data offer us unprecedented opportunities to characterize functional roles of genetic variants that underlie human complex phenotypes, such as height, weight, blood pressure and disease status. To establish the causal link from genotypes to organismal phenotypes, there is a great need to perform integrative analysis of multilayered annotation data.

A big challenge in integrative analysis is how to put multilayered information into a unified model and automatically select most relevant genomic features from a potentially huge set of genomic features. In this talk, we introduce a flexible statistical approach, named IMAC, to integrating multilayered annotation for characterizing functional roles of genetic variants that underlie human complex phenotypes. IMAC enabled us to automatically perform feature selection from a large number of annotated genomic features and naturally incorporate the selected features for prioritization of genetic risk variants. IMAC not only demonstrated a remarkably computational efficiency (e.g., it took about 2$\sim$3 minutes to handle millions of genetic variants and thousands of functional annotations), but also allowed rigorous statistical inference of the model parameters and false discovery rate control in risk variant prioritization. With the IMAC approach, we performed integrative analysis of genome-wide association studies on multiple complex human traits and genome-wide annotation resources, e.g., expression QTL and splicing QTL. The analysis results revealed interesting regulatory patterns of risk variants. These findings undoubtedly deepen our understanding of genetic architectures of complex traits. The underlying statistical principle in IMAC design is fairly general, the key idea can be leveraged to other Big Data involved applications.

High Order Bias Corrected Estimator for Time-Average Variance Constant

Chun Yip Yau, The Chinese University of Hong Kong

Estimation of time-average variance constant (TAVC), which is the asymptotic variance of the sample mean of a time series, is of fundamental importance in statistical inference. In this paper, by considering high order corrections to the asymptotic biases, we develop a new class of TAVC estimators that enjoys optimal $\mathcal{L}^2$- convergence rates under different strength of the serial dependence of the time series. Comparisons to existing TAVC estimators are comprehensively investigated. In particular, the proposed high order corrected estimator has the best performance in terms of mean squared error.

Augmenting the Unreturned for Field Data with Information on Returned Failures Only

Zhisheng Ye, National University of Singapore

Field data are an important source of reliability information for many commercial products. Because field data are often collected by the maintenance department, information on failed and returned units is well maintained. Nevertheless, information on unreturned units is generally unavailable. The unavailability leads to truncation in the lifetime data. This study proposes a data augmentation algorithm for this type of truncated field return data with returned failures available only.

The algorithm is based on an idea to reveal the hidden unobserved lifetimes. Theoretical justifications of the procedure for augmenting the hidden unobserved are given. On the other hand, the algorithm is iterative in nature. Asymptotic properties of the estimators from the iterations are investigated. Both point estimation and the information matrix of the parameters can be directly obtained from the algorithm. In addition, a by-product of the algorithm is a non-parametric estimator of the installation time distribution. An example from an asset-rich company is given to demonstrate the proposed methods.

Online Pairwise Learning Algorithms (OPERA)

Yiming Ying, University at Albany, State University of New York

---

Pairwise learning usually refers to a learning task which involves a loss function depending on pairs of examples, among which most notable ones include bipartite ranking, metric learning and AUC maximization. The main challenge in pairwise learning is the "bigger" volume of pairs of examples in the sense that the number of pairs of examples grows quadratically wrt the number of examples. Stochastic online learning is widely used to handle such large-scale and fast-updating data. In this presentation, I will talk about our recent work on analyzing the convergence of stochastic online learning algorithms for pairwise learning in RKHSs, which we refer to as the Online Pairwise lEaRning Algorithm (OPERA). Specifically, we establish the almost surely convergence for the last iterate of OPERA without any assumptions on the underlying distribution. Explicit convergence rates are derived under the condition of polynomially decaying step sizes. Our analysis mainly depends on the characterization of RKHSs using its associated integral operators and probability inequalities for random variables with values in a Hilbert space.

Big Data Analytics in Online Education

Aijun Zhang, Hong Kong Baptist University

Online educational systems generate large amounts of real-time streaming data, especially after 2012 the year of the MOOC. Many of recent innovations in big data can be adopted to develop learning analytics for online education systems. In this talk we will discuss two big data applications in online education, including the dropout prediction in an MOOC platform and the two-way scoring models in online testing and assessment.

An Extended Single Index Model with Missing Response at Random

Tao Zhang, Guangxi University of Science and Technology

An extended single-index model is considered when responses are missing at random. A three-step estimation procedure is developed to define an estimator for the single index parameter vector by a joint estimating equation. The proposed estimator is shown to be asymptotically normal. An iterative scheme for computing this estimator is proposed. This algorithm only involves one-dimensional nonparametric smoothers, thereby avoiding the data sparsity problem caused by high model dimensionality. Some simulation study is conducted to investigate the finite sample performances of the proposed estimators.

Globally Efficient Nonparametric Inference of Average Treatment Effects by Empirical Balancing Calibration Weighting

Zheng Zhang, The Chinese University of Hong Kong

The estimation of average treatment effects based on observational data is extremely important in practice and has been studied by generations of statisticians under different frameworks. Existing globally efficient estimators require non-parametric estimation of a propensity score function, an outcome regression function or both, but their performance can be poor in practical sample sizes. Without explicitly estimating either function, we consider a wide class calibration weights constructed to attain an exact three-way balance of the moments of observed covariates among the treated, the controls, and the combined group. The wide class includes exponential tilting, empirical likelihood and generalized regression as important special cases, and extends survey calibration estimators to different statistical problems and with important distinctions. Global semiparametric efficiency for the estimation of average treatment effects is established for this general class of calibration estimators. The results show that efficiency can be achieved by solely balancing the covariate distributions without resorting to direct estimation of propensity score or outcome regression. We also propose a consistent estimator of the efficient asymptotic variance, which does not involve additional functional estimation of either the propensity score or the outcome regression functions. The proposed variance estimator outperforms existing estimators that require a direct approximation of the efficient influence function.

Joint work with Gary Chan and Phillip Yam.