Program Schedule

Venue: Lo Ka Chung Building, Lee Shau Kee Campus of HKUST

19 December 2013 (Thursday)		
Time	*Event*	*Event*
8:30 – 9:00	Registration (Venue: IAS Lecture Theater)	
9:00 – 9:10	Welcome Remarks by Nancy Ip *[HKUST]*	
9:10 – 10:10	Talk #1: "Quality Technology in the High-Tech Age" by Jeff Wu *[Georgia Institute of Technology]*	
10:10 – 10:30	Refreshments	
	(Venue: 2/F Seminar Room)	(Venue: 4/F Seminar Room)
	Session 1 – Chair: Bingyi Jing	Session 2 – Chair: Xianhua Peng
10:30 – 11:00	Talk #2: "Sample Size Requirement For Some Low-Dimensional Estimation Problems" by Cun-Hui Zhang *[Rutgers, The State University of New Jersey]*	Talk #3: "New Test for One-Way ANOVA with Functional Data and Application to Ischemic Heart Screening" by Ming-Yen Cheng *[National Taiwan University]*
11:00 – 11:30	Talk #4: "A Sequential Penalized Likelihood Approach for Feature Selection with Ultra-High Dimensional Feature Space" by Zehua Chen *[National University of Singapore]*	Talk #5: "Group Regularized Estimation under Strong Hierarchy" by Yiyuan She *[Florida State University]*
11:30 – 11:45	Tea Break	
	Session 3 – Chair: Shujie Ma	Session 4 – Chair: Fugee Tsung
11:45 – 12:15	Talk #6: "Imprinting Test of Disease-Associated SNPs under Mixture Model" by Jianhua Guo *[Northeast Normal University]*	Talk #7: "Monitoring High-Dimensional Data Streams" by Wei Jiang *[Shanghai Jiao Tong University]*
12:15 – 12:45	Talk #8: "Statistical Inference of Covariate-Adaptive Randomized Clinical Trials" by Feifang Hu *[George Washington University]*	Talk #9: "BIG Data and Dimensional Analysis" by Dennis Lin *[Pennsylvania State University]*

12:45 – 14:00	Lunch Break	
	Session 5 – Chair: Lanceclot James	Session 6 – Chair: Shiqing Ling
14:00 – 14:30	Talk #10: "Estimation and Selection of Complex Effects in Pooled Nested Case-Control Studies with Heterogeneity" by Wenbin Lu *[North Carolina State University]*	Talk #11: "Interval Estimation for Random Coefficient AR model and Predictive Regressions" by Liang Peng *[Georgia Institute of Technology]*
14:30 – 15:00	Talk #12: "Semi Parametric Modeling for Nonlinear Interactions" by Shujie Ma *[University of California, Riverside]*	Talk #13: "Optimizing Organ Exchange System – Celebrating the 2012 Nobel Prize in Economics" by Peter Song *[University of Michigan]*
15:00 – 15:15	Tea Break	
	Session 7 – Chair: Inchi Hu	Session 8 – Chair: Yingying Li
15:15 – 15:45	Talk #14: "A Partially Linear Single-Index Transformation Model" by Qihua Wang *[Chinese Academy of Sciences]*	Talk #15: "Estimation of the Error Correlation Matrix in Semi-Parametric Models for Brain fMRI Data" by Chunming Zhang *[University of Wisconsin – Madison]*
15:45 – 16:15	Talk #16: "Oracally Efficient Estimation of Autoregressive Error Distribution with Simultaneous Confidence Band" by Lijian Yang *[Michigan State University]*	Talk #17: "On Consistency of Community Detection in Networks" by Ji Zhu *[University of Michigan]*
16:15 – 16:30	Refreshments	
	Session 9 – Chair: Kani Chen	Session 10 – Chair: Xinghua Zheng
16:30 – 17:00	Talk #18: "Generalized Measures of Correlation and Their Implications in GARCH and Heston Models" by Zhengjun Zhang *[University of Wisconsin – Madison]*	Talk #19: "Adaptive Bayesian Estimation" by Harrison Huibin Zhou *[Yale University]*
17:00 – 17:30	Talk #20: "Semiparametric Estimation of Treatment Effect with Logistic Regression Model" by Yong Zhou *[Chinese Academy of Sciences and Shanghai University of Finance and Economics]*	Talk #21: "Sparse PCA: Optimal Rates and Adaptive Estimation" by Tony Cai *[University of Pennsylvania]*

Abstracts

Quality Technology in the High-Tech Age

(Talk #1)

Jeff Wu

Georgia Institute of Technology

Traditional work in quality control and improvement had its main impetus from needs in manufacturing like autos, electronics and chemicals. As the economies in developed countries are increasingly dependent on high value-added products, it is natural to ask what role can quality professionals play in the high-tech age. Without the benefit of crystal gazing, I will use several research examples to search for hints to the future. I would describe the essence of some research projects, including the study of nano-mechanical properties, computer-aided design and management of data centers, and electronics packaging. The following features distinguish these problems from traditional quality technology. The experiments are more elaborate and the responses can be sensitive to input conditions; they often require complex modeling; how to scale up from lab conditions to industrial productions; they often use extensive computer modeling (e.g., finite element analysis) before conducting (or even foregoing) physical experiments for verification. Based on these new features, I will try to address the ultimate question: “will there be a paradigm shift and what to expect”?

Sample Size Requirement for Some Low-Dimensional Estimation Problems

(Talk #2)

Cun-Hui Zhang

Rutgers, The State University of New Jersey

We consider sample size requirement for estimating low-dimensional parameters at the parametric rate with high-dimensional data. For the estimation of elements of a precision matrix, we provide a sufficient and necessary condition on the sample size under a side condition on the range of the spectrum of the precision matrix. Efficient estimators are provided in a semiparametric sense when the sample size is sufficiently large but still allowed to be of smaller order than the size of the precision matrix. However, the sample size requirement can be substantially weakened in certain associated regression problems. This is based on joint work with Zhao Ren, Tingni Sun, Stephanie Zhang, and Harrison Zhou.

New Test for One-Way ANOVA with Functional Data and Application to Ischemic Heart Screening

(Talk #3)

Ming-Yen Cheng

National Taiwan University

We propose and study a new global test for the one-way ANOVA problem in functional data analysis. The test statistic is taken as the maximum value of the usual pointwise F-test statistics over the interval the functional responses are observed. A nonparametric bootstrap method is employed to approximate the null distribution of the test statistic and to obtain an estimated critical value for the test. The asymptotic random expression of the test statistic is derived and the asymptotic power is studied. In particular, under mild conditions, the proposed test asymptotically has the correct level and is root-n consistent in detecting local alternatives. Via some simulation studies, it is found that in terms of both level accuracy and power, the proposed test outperforms the Globalized Pointwise F test of Zhang and Liang (2013) when the functional data are highly or moderately correlated, and its performance is comparable with the latter otherwise. An application to an ischemic heart real dataset suggests that, after proper manipulation, resting electrocardiogram (ECG) signals can be used as an effective tool in clinical ischemic heart screening, without the need of further stress tests as in the current standard procedure.

A Sequential Penalized Likelihood Approach for Feature Selection with Ultra-High Dimensional Feature Space

(Talk #4)

Zehua Chen

National University of Singapore

In this talk, we discuss a sequential method for variable selection in ultra-high dimensional feature space — the sequential penalized likelihood cum EBIC approach. In principle, the method selects features by sequentially solving partially penalized likelihood problems where the features selected in earlier steps are not penalized and uses the extended BIC (EBIC) as the stopping rule. We first consider the method in the case of linear models. The asymptotic properties of the method are considered under the setting that the dimension of the feature space is ultra-high and the number of relevant feature diverges. We show that, with probability converging to 1, the method first selects all the relevant features before any irrelevant features can be selected, and that the EBIC decreases until it attains the minimum at the model consisting of exactly all the relevant features and then begins to increase. These results establish the selection consistency of the method. Then we consider the generation of the method to the case of generalized linear models. We also consider the method for models with interactions. The sequential penalized likelihood cum EBIC approach is compared with other methods by simulation studies, which demonstrates its edge over the other methods for feature selection in ultra-high dimensional space.

Group Regularized Estimation under Strong Hierarchy

(Talk #5)

Yiyuan She

Florida State University

In high-dimensional models involving between-term interactions, statisticians usually favor variable selection obeying certain logical hierarchical constraints. The talk focuses on strong hierarchy which means that the existence of an interaction term implies that both associated main effects must be present. Although lately the hierarchical lasso has been proposed, the existing computational algorithms converge very slow and cannot meet the challenge of big data. There is a lack of finite-sample studies in the literature partially due to the difficulty that multiple sparsity-promoting penalties are enforced on the same subject. We propose a new estimator based on group multi-regularization to capture various types of structural parsimony. We present some nonasymptotic results and develop a general-purpose algorithm with a theoretical guarantee of strict iterate convergence. A predictive information criterion is proposed for data-dependent tuning and can help achieve the optimal error rate in a minimax sense.

Imprinting Test of Disease-Associated SNPs under Mixture Model

(Talk #6)

Jianhua Guo

Northeast Normal University

Genomic imprinting represents a known aspect of the etiology of schizophrenia, a serious and common neuropsychiatric disease. This study delineates the role of imprinting in the relationships between schizophrenia-associated single nucleotide polymorphisms (SNPs) of the GABRB2 gene for its beta2 subunit of $GABA_A$ receptors and the quantitative trait of GABRB2 mRNA expression. The imprinting phenomenon depicts differential expression levels of the allele depending on its parental origin. When the parental origin is unknown, the expression level has a finite normal mixture distribution. In such application, a random sample on expression levels consists of three subsamples according to the number of minor alleles an individual possesses, one of which is the mixture and the rest two are homogeneous. This understanding leads to a likelihood ratio test (LRT) for the presence of the imprinting. Due to non-regularity of the finite mixture model, the classical asymptotic conclusions on likelihood-based inferences are inapplicable. We show that the maximum likelihood estimator of the mixing distribution remains consistent. More interestingly, helped by homogeneous subsamples, the LRT statistic has an elegant and rather distinct mixture chi-squared null limiting distribution. Simulation studies confirm that the limiting distribution provides precise approximations to the finite sample distributions in various parameter settings. The LRT is applied to expression data sets on the schizophrenia susceptibility gene GABRB2. Our analyses provide evidences of imprintings on a number of isoform expressions of its $GABA_A$ receptor beta2 subunit protein encoded by the GABRB2 gene.

Monitoring High-Dimensional Data Streams

(Talk #7)

Wei Jiang

Shanghai Jiao Tong University

Monitoring high-dimensional data streams has become increasingly important for real-time detection of abnormal activities in many data-rich applications. We are interested in detecting an occurring event as soon as possible, but we do not know which subset of data streams is affected by the event. By connecting to the problem of detecting heterogeneous mixtures, a new control chart scheme is developed based on a powerful goodness-of-fit test of the local cumulative sum statistics from each data stream. Both (asymptotically) theoretical analysis and numerical results show that the proposed method is able to balance the detection of various fractions of affected streams, and generally outperforms existing methods.

Statistical Inference of Covariate-Adaptive Randomized Clinical Trials

(Talk #8)

Feifang Hu

George Washington University

In a short period of time, advances in genetics have allowed scientists to identify genes (biomarkers) that are linked with certain diseases. To translate these great scientific findings into real-world products for those who need them, clinical trials play an essential and important role. To develop personalized medicine, covariate-adaptive randomized clinical trials should be implemented so that genetics information and other biomarkers can be incorporated in treatment selection. However, the properties of conventional statistical methods are not well studied in literature. In this talk, theoretical properties of conventional statistical methods are derived under general covariate-adaptive randomized clinical trials. Some further statistical issues are also discussed.

BIG Data and Dimensional Analysis

(Talk #9)

Dennis Lin

Pennsylvania State University

The ability to parse information, faster and deeper, is allowing researchers to understand the world in a way they could only dream about before. In this talk (first portion), the common understanding about BIG data will be discussed. Its challenges and impacts to Statistics will be addressed. In particular, we are interested in what types of Statistics are needed for BIG data era. On the other hand, Dimensional Analysis (DA) is a fundamental method in the engineering and physical sciences for analytically reducing the number of experimental variables prior to the experimentation. The method is of great generality. In this talk (second portion), an overview/introduction of DA will be given. Some initial ideas on using DA for BIG data will be discussed. Future research issues will be proposed.

Estimation and Selection of Complex Effects in Pooled Nested Case-Control Studies with Heterogeneity

(Talk #10)

Wenbin Lu

North Carolina State University

A major challenge in cancer epidemiologic studies, especially those of rare cancers, is observing enough cases. To address this, researchers often join forces by bringing multiple studies together to achieve large sample sizes, allowing for increased power in hypothesis testing, and improved efficiency in effect estimation. Combining studies, however, renders the analysis difficult owing to the presence of heterogeneity in the pooled data. In this work, we propose an adaptive group lasso (gLASSO) penalized likelihood approach to simultaneously identify important variables and estimate their effects, and a composite agLASSO penalized approach to identify variables with heterogeneous effects. Simulation studies and an application to the pooled ovarian cancer study will be presented to illustrate the performance of our proposed approaches.

Interval Estimation for Random Coefficient AR Model and Predictive Regressions

(Talk #11)

Liang Peng

Georgia Institute of Technology

The quasi maximum likelihood estimation of random coefficient autoregressive models requires coefficient randomness if non-stationary cases are allowed. In the first half of the talk, we propose empirical likelihood methods based on a weighted score equation to construct a confidence interval for the coefficient. We do not need to distinguish whether the coefficient is random or deterministic and whether the process is stationary or non- stationary. A simulation study confirms the good finite sample behavior of the proposed methods, and we apply our methods to study U.S. macroeconomic data.

In the second half of the talk, we study the predictive regression models in financial econometrics. For testing the predictability, procedures based on the least squares estimator and other bias-corrected estimators proposed in the literature suffer from a complicated asymptotic limit, which depends on whether the predicting variable is stationary or non-stationary or has an infinity variance.

In this paper, we propose a novel weighted estimation approach to estimate the coefficient, which always has a normal limit regardless of predicting variable being stationary, or non-stationary or having an infinite variance. However, the asymptotic variance still depends on the case of stationary or non-stationary. To overcome this difficulty, we propose a unified (empirical likelihood) method to test the predictability without estimating the asymptotic variance.

Semi Parametric Modeling for Nonlinear Interactions

(Talk #12)

Shujie Ma

University of California, Riverside

It has been a long history of utilizing interactions in regression analysis to investigate alterations in covariate-effects on response variables. In this paper we aim to address two kinds of new challenges resulted from the inclusion of such high-order effects in the regression model for complex data. The first kind arises from a situation where interaction effects of individual covariates are weak but those of combined covariates are strong, and the other kind pertains to the presence of nonlinear interactive effects directed by low-effect covariates. We propose a new class of semiparametric models with varying index coefficients, which enables us to model and assess nonlinear interaction effects between grouped covariates on the response variable. As a result, most of the existing semiparametric regression models are special cases of our proposed models. We develop a numerically stable and computationally fast estimation procedure utilizing both profile least squares method and local fitting. We establish both estimation consistency and asymptotic normality for the proposed estimators of index coefficients as well as the oracle property for the nonparametric function estimator.

Optimizing Organ Exchange System – Celebrating the 2012 Nobel Prize in Economics

(Talk #13)

Peter Song

University of Michigan

The 2012 Nobel Prize in Economics is awarded to Alvin E. Roth and Lloyd S. Shapley for the theory of stable allocations and the practice of market design. The most influential work made by Alvin Roth was his seminal theory of matching organ donors with patients. This novel empirical study has provided and enhanced Shapley's basic theory and the application of Gale-Shapley algorithm in medical practice of organ exchanges. In this talk, I will first present an overview of Alvin Roth's classical static organ exchange system, including graphic formulation and integer programming optimization. I will then introduce our recent work on probabilistic organ exchange systems and optimal strategies of matching organs from living donors with biologically unrelated recipients. Through various microsimulation models, I will show that our new organ donation system outperforms Roth's organ donation system. This is a joint work with Jack Kalbfleisch, John Li, Yanhua Chen, Yan Zhou, Alan Leitchman, and Michael Rees.

A Partially Linear Single-Index Transformation Model

(Talk #14)

Qihua Wang

Chinese Academy of Sciences

We propose a constrained least square estimator of the transformation function of a partially linear single-index transformation model, where the transformation function, single-index function and error distribution are all nonparametric. The estimators of the regression coefficients and the single-index function are provided by the similar idea to the minimum average variance estimation method. Basis function approximation is employed to estimate the transformation and single-index functions, and cross validation criteria are proposed to select suitable sets of basis functions. Asymptotical properties of the estimators in the sense of almost sure convergence are established. Simulations studies show that our proposed estimators work well. A real-world data analysis of total health care charges was used to illustrate the proposed procedure.

Estimation of the Error Correlation Matrix in Semi-Parametric Models for Brain fMRI Data

(Talk #15)

Chunming Zhang

University of Wisconsin – Madison

In statistical analysis of functional magnetic resonance imaging (fMRI), dealing with the temporal correlation is a major challenge in assessing changes within voxels. In this paper, we aim to address this issue by considering a semi-parametric model for fMRI data. For the error process in the semi-parametric model, we construct a banded estimate of the auto-correlation matrix R, and propose a refined estimate of the inverse of R. Under some mild regularity conditions, we establish consistency of the banded estimate with an explicit convergence rate and show that the refined estimate converges under an appropriate norm. Numerical results suggest that the refined estimate performs conceivably well when it is applied to the detection of the brain activity.

Oracally Efficient Estimation of Autoregressive Error Distribution with Simultaneous Confidence Band

(Talk #16)

Lijian Yang

Michigan State University

We propose kernel estimator for the distribution function of unobserved errors in autoregressive time series, based on residuals computed by estimating the autoregressive coefficients with Yule-Walker method. Under mild assumptions, we establish oracle efficiency of the proposed estimator, i.e., it is asymptotically as efficient as the kernel estimator of the distribution function based on the unobserved error sequence itself. Applying result of \cite{WCY13}, the proposed estimator is also asymptotically indistinguishable from the empirical distribution function based on the unobserved errors. A smooth simultaneous confidence band (SCB) is then constructed based on the proposed smooth distribution estimator and Kolmogorov distribution. Simulation examples support the asymptotic theory.

On Consistency of Community Detection in Networks

(Talk #17)

Ji Zhu

University of Michigan

Community detection is a fundamental problem in network analysis, with applications in many diverse areas. The stochastic block model is a common tool for model-based community detection, and asymptotic tools for checking consistency of community detection under the block model have been recently developed by Bickel and Chen (2009).

However, the block model is limited by its assumption that all nodes within a community are stochastically equivalent, and provides a poor fit to networks with hubs or highly varying node degrees within communities, which are common in practice. The degree-corrected block model (Karrer and Newman 2010) was proposed to address this shortcoming, and allows variation in node degrees within a community while preserving the overall block model community structure. In this paper, we establish general theory for checking consistency of community detection under the degree-corrected block model, and compare several community detection criteria under both the standard and the degree-corrected block models. We show which criteria are consistent under which models and constraints, as well as compare their relative performance in practice. We find that methods based on the degree-corrected block model, which includes the standard block model as a special case, are consistent under a wider class of models; and that modularity-type methods require parameter constraints for consistency, whereas likelihood-based methods do not. On the other hand, in practice the degree correction involves estimating many more parameters, and empirically we find it is only worth doing if the node degrees within communities are indeed highly variable. We illustrate the methods on simulated networks and on a network of political blogs. This is joint work with Yunpeng Zhao and Elizaveta Levina.

Generalized Measures of Correlation and Their Implications in GARCH and Heston Models

(Talk #18)

Zhengjun Zhang

University of Wisconsin – Madison

Applicability of Pearson's correlation as a measure of explained variance is by now well understood. One of its limitations is that it does not account for asymmetry in explained variance. Aiming to obtain broad applicable correlation measures, we use a pair of r-squares of generalized regression to deal with asymmetries in explained variances, and linear or nonlinear relations between random variables.

We call the pair of r-squares of generalized regression generalized measures of correlation (GMC). We present examples under which the paired measures are identical, and they become a symmetric correlation measure which is the same as the squared Pearson's correlation coefficient. As a result, Pearson's correlation is a special case of GMC. Theoretical properties of GMC show that GMC can be applicable in numerous applications and can lead to more meaningful conclusions and decision making. In statistical inferences, the joint asymptotics of the kernel based estimators for GMC are derived and are used to test whether or not two random variables are symmetric in explaining variances. The testing results give important guidance in practical model selection problems. In real data analysis, this talk presents ideas of using GMCs as an indicator of suitability of asset pricing models, and hence new pricing models may be motivated from this indicator.

Adaptive Bayesian Estimation

(Talk #19)

Harrison Huibin Zhou

Yale University

A block prior is proposed for adaptive Bayesian estimation of various nonparametric models. The posterior contraction rate is shown to be optimal adaptively for a large collection of Sobolev balls.

If time permits, adaptive Bayesian estimation for sparse PCA will be discussed.

Coauthor: Chao Gao

Semiparametric Estimation of Treatment Effect with Logistic Regression Model

(Talk #20)

Yong Zhou

Chinese Academy of Sciences and
Shanghai University of Finance and Economics

Treatment effect is an important index in comparing two-sample data in survival analysis, industry manufacture, clinical medicine and many other applications. In this paper, we propose a unified semiparametric approach to estimate different types of treatment effects under a case-control sampling plan with the logistic regression model assumption, which is equivalent to a two-sample density ratio model. For different treatment effects, we construct different estimating functions and the nuisance parameters in estimating functions are estimated firstly by the empirical likelihood method. Here, we allow that the functions are nonsmooth with respect to parameters. The confidence interval for the treatment effect based on the empirical likelihood ratio method is also presented. We prove that the estimator based on the estimating equation is consistent and asymptotically normal and the empirical log-likelihood ratio statistic has a limiting scaled chi-square distribution. Simulation studies are reported to assess the finite sample properties of the proposed estimator and the performance of the confidence interval. The proposed methods are applied to real data examples and some interesting results are presented.

This is joint work with Cunjie Lin, and Wenhua Wei, Academy of Mathematics and Systems Science, Chinese Academy of Sciences.

Sparse PCA: Optimal Rates and Adaptive Estimation

(Talk #21)

Tony Cai

University of Pennsylvania

Principal component analysis (PCA) is one of the most commonly used statistical procedures with a wide range of applications. In this talk we consider both minimax and adaptive estimation of the principal subspace in the high dimensional setting. The optimal rates of convergence are established for estimating the principal subspace which are sharp with respect to all the parameters, thus providing a complete characterization of the difficulty of the estimation problem in term of the convergence rate. We then introduce an adaptive procedure for estimating the principal subspace which is fully data driven and can be computed efficiently. It is shown that the estimator attains the optimal rates of convergence simultaneously over a large collection of the parameter spaces. A key idea in our construction is a reduction scheme which reduces the sparse PCA problem to a high-dimensional multivariate regression problem. This method is potentially also useful for other related problems. This is joint work with Zongming Ma and Yihong Wu.