Workshop on Statistical theories, methodologies and applications in high dimensional data analysis and biomedicine

March 19

Room 1217, Department of Mathematics, Hong Kong Baptist University

09:30-09:40 Welcoming Remarks

9:40-10:20 Prof. Jeff Yao, Department of Statistics and Actuarial Science, The University of Hong Kong

Title: Identifying the number of factors from singular values of a large sample auto-covariance matrix

Abstract: Identifying the number of factors in a high-dimensional factor model has attracted much attention in recent years and a general solution to the problem is still lacking. A promising ratio estimator based on singular values of lagged sample autocovariance matrices has been recently proposed in the literature with a reasonably good performance under some specific assumption on the strength of the factors. Inspired by this ratio estimator and as a first main contribution, this paper proposes a complete theory of such sample singular values for both the factor part and the noise part under the large-dimensional scheme where the dimension and the sample size proportionally grow to infinity. In particular, we provide an exact description of the phase transition phenomenon that determines whether a factor is strong enough to be detected with the observed sample singular values. Based on these findings and as a second main contribution of the paper, we propose a new estimator of the number of factors which is strongly consistent for the detection of all significant factors (which are the only theoretically detectable ones). In particular, factors are assumed to have the minimum strength above the phase transition boundary which is of the order of a constant; they are thus not required to grow to infinity together with the dimension (as assumed in most of the existing papers on high-dimensional factor models). Empirical Monte-Carlo study as well as the analysis of stock returns data attest a very good performance of the proposed estimator. In all the tested cases, the new estimator largely outperforms the existing estimator using the same ratios of singular values. This is a joint work with Ms. Zeng Li and Ms. Qinwen Wang (The University of Hong Kong).

10:20 – 10: 40 Prof. Liping Zhu, Institute of Statistics and Big Data, Renmin University

Title: A post-screening diagnostic study in sufficient dimension reduction for ultrahigh dimensional data

Abstract: In this talk I will introduce a consistent lack-of-fit test to examine whether or not replacing the original ultrahigh dimensional covariates with a given number of linear combinations will result in loss of regression information. To attenuate spurious correlations which are often seen in ultrahigh dimensional covariates and may substantially inflate type-I error rates, we suggest to randomly split the observations into two halves. In the first halve of observations we screen out as many irrelevant covariates as possible. This helps us reduce the ultrahigh dimensionality to a moderate scale. In the second halve we perform a lack-of-fit test for conditional independence within the context of sufficient dimension reduction. This data-splitting strategy helps us retain the type-I error rate pretty well. We propose a new statistic to test conditional independence, and show that our propose test procedure is n-consistent under the null and root-n-consistent under the alternative hypothesis. Our proposed test procedure is consistent in the sense that it has nontrivial power against all feasible alternatives. In addition, we suggest a bootstrap procedure to decide critical values and show that our

bootstrap procedure is consistent. We demonstrate the effectiveness of our test procedure through comprehensive simulations and an application to the rats' red-eye data set.

10:50 – 11:30 Heng Peng, Department of Mathematics, HKBU.

Title: Model Selection for Gaussian Mixture Models

Abstract: This talk is concerned with an important issue in finite mixture modeling, namely the selection of the number of mixing components. A new penalized likelihood method is proposed for finite multivariate Gaussian mixture models, and it is shown to be statistically consistent in determining the number of components. A modified EM algorithm is developed to simultaneously select the number of components and estimate the mixing probabilities and the unknown parameters of Gaussian distributions. Simulations and a real data analysis are presented to illustrate the performance of the proposed method.

11: 30 – 13:30 Lunch break

13:30 – 14:10 Xiaodan Fan, Department of Statistics, The Chinese University of Hong Kong

Title: An HMM model for the identification of differentially methylated regions

Abstract: Methylation is one of the most informative epigenetic modifications that are currently widely studied. Nowadays, disease studies are producing tons of genome-wide methylation data from case-control design. Many efforts have been devoted to detecting differentially methylated regions (DMRs) for biological inference. DMR is the region where multiple adjacent CpG sites show different methylation status between phenotypes, which may occur throughout the whole genome. In this work, we propose to use Non-homogeneous Hidden Markov Model for methylation modeling and DMR detection. We proposed several combinations of different transition models and response models, and evaluated them by both synthetic data and real data. Our methods outperformed existing methods.

14:10-14:50 Minghua Deng, School of Mathematical Sciences, Peking University

Title: CCLasso: correlation inference for compositional data through Lasso

Abstract: Direct analysis of microbial communities in the environment and human body has become more convenient and reliable owing to the advancements of high-throughput sequencing techniques for 16S rRNA gene profiling. Inferring the correlation relationship among members of microbial communities is of fundamental importance for genomic survey study. Traditional Pearson correlation analysis treating the observed data as absolute abundances of the microbes may lead to spurious results because the data only represent relative abundances. Special care and appropriate methods are required prior to correlation analysis for these compositional data. In this paper, we first discuss the correlation definition of latent variables for compositional data. We then propose a novel method called CCLasso based on least squares with l1 penalty to infer the correlation network for latent variables of compositional data from metagenomic data. An effective alternating direction algorithm from augmented Lagrangian method is used to solve the optimization problem. The simulation results show that CCLasso outperforms existing

methods, e.g. SparCC, in edge recovery for compositional data. It also compares well with SparCC in estimating correlation network of microbe species from the Human Microbiome Project.

14:50 – 15:10 Coffee Break

15:10 – 15:50 Rui Jiang, Statistics and Bioinformatics, Tsing Hua University

Title: Identification of disease-causing single nucleotide variants in exome sequencing studies

Abstract: Exome sequencing has been widely used in detecting pathogenic nonsynonymous single nucleotide variants (SNVs) for human inherited diseases. However, traditional statistical genetics methods are ineffective in analyzing exome sequencing data, due to such facts as the large number of sequenced variants, the presence of non-negligible fraction of pathogenic rare variants or de novo mutations, and the limited size of affected and normal populations. Here, we propose bioinformatics approaches, SPRING, snvForest and GLINTS, for identifying pathogenic nonsynonymous SNVs for a given query disease. SPRING integrates six functional effect scores calculated by existing methods and five association scores derived from a variety of genomic data sources to calculate the statistical significance that an SNV is causative for a query disease. snvForest adopts an ensemble learning method to assign prediction scores to candidate SNVs. These methods are designed to use with a set of seed genes known as associated with the disease of interest, and thus is suitable for studies on diseases with some prior knowledge. GLINTS further incorporates three disease phenotype similarity data to facilitate the detection of causative SNVs without any knowledge of seed genes for a query disease. This method is therefore suitable for research on diseases whose genetic bases are completely unknown. With a series of comprehensive validation experiments, we demonstrate the effectiveness of these methods, not only in simulation studies, but also in detecting causative de novo mutations for autism, epileptic encephalopathies and intellectual disability.

15:50-16:30 Yingying Wei, Department of Statistics, The Chinese University of Hong Kong
Title: Are all transcription factors interacting with each other?

Abstract: Understanding the interactions of different transcription factors (TF)s is a crucial first step toward deciphering gene regulatory mechanism. With advances of high-throughput sequencing technology such as ChIP-seq, the genome-wide binding sites of many TFs have been profiled under different biological contexts. It is of great interest to quantify the interactions among different TFs. Analyses of the overlapping patterns of binding sites have been widely performed, mostly based on ad hoc methods. Due to the heterogeneity and the tremendous size of the genome, such methods often lead to biased even erroneous results. In the first part of the talk, we introduce a Simpson's paradox type of phenomenon in assessing the genome-wide spatial correlation of TF binding sites. Leveraging information from publicly available data, we propose a testing procedure for evaluating the significance of overlapping from a pair of proteins, which accounts for background artifacts and genome heterogeneity. In the second part of the talk, we propose a dynamic Poisson graphic model to characterize the co-activation patterns among TFs across the genome and under diverse biological conditions.

16:30 – 17:10 Lin Hou, Center for Statistical Science, Tsinghua University

Title: The impact of genotyping errors on the statistical power of association test in genome-wide association studies.

Abstract: A key step in genome-wide association studies (GWAS) is to infer genotypes across millions of markers from high throughput measurements for each individual, either using microarrays or sequencing. Accurate genotype calling is essential for downstream statistical analysis of phenotype genotype associations. Recently, next generation sequencing (NGS) has become more commonly used for genotyping in GWAS, due to reduced cost and ability to study a more complete spectrum of markers, including rare variants and unknown variants. Various pipelines have been developed for variant/genotype calling from NGS data, and the accuracy of variant calling has been extensively evaluated. However, how variant calling accuracy affects downstream association analysis has not been studied based on empirical data where both microarrays and NGS were used for inferring genotypes. In this article, we investigate the impact of variant calling errors on the statistical power to identify association between disease and single nucleotides and that between disease and multiple rare variants. Our results show that the power of burden test for rare variants is strongly influenced by the specificity in variant calling, but rather robust to sensitivity.