Workshop on Data Science in Biomedicine

July 6

Room 1217, Department of Mathematics, Hong Kong Baptist University

09:30-09:40 Welcoming Remarks

9:40-10:20 Pak Chung Sham, Centre for Genomic Sciences, The University of Hong Kong

Title: Characterizing polygenes from GWAS data

Abstract: The results of genome-wide association studies (GWAS) indicate that most complex diseases are highly polygenic: (1) associated alleles typically have small effect sizes, (2) the number of significant associated alleles has continued to increase with increasing sample size, and (3) genome-wide methods such as random effects models have detected much greater total genetic effects than accounted for by genome-wide significant SNPs. The genome-wide significant loci so far detected for complex diseases therefore likely represent the tip of the iceberg of the total polygenic components of these diseases. New statistical methods, informed by genome organization and function, will be necessary to characterize the polygenic components of complex diseases. Examples of such methods will be presented.

10:10 – 10: 40 Yingying Wei, Statistics, The Chinese University of Hong Kong

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

Abstract: 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-linespecific open chromatin and DNA methylation landscape map. This is a joint work with Mai Shi.

10:40 – 11: 00 Break

11:00 – 11: 30 Can Yang

Title: IPAC: A Flexible Statistical Approach to Integrating Pleitoropy and Annotation for Characterizing Functional Roles of Genetic Variants that Underlie Human Complex Phenotypes

Abstract: 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, e.g., epigenome and transcriptome. On the other hand, increasing evidence suggests that seemly unrelated phenotypes can share common genetic factors, which is known as pleiotropy. A big challenge in integrative analysis is how to put pleiotropy and annotation 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 IPAC, to integrating pleiotropy and annotation for characterizing functional roles of genetic variants that underlie human complex phenotypes. IPAC 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. IPAC not only demonstrated a remarkably computational efficiency (e.g., it took about 2~3 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 IPAC approach, we performed integrative analysis of genome-wide association studies on multiple complex human traits and genome-wide annotation resources, e.g., Roadmap epigenome. The analysis results revealed interesting regulatory patterns of risk variants. These findings undoubtedly deepen our understanding of genetic architectures of complex traits. This is a joint work with Dongjun Chung, Cong Li, Jin Liu, Xiang Wan, Qian Wang, Chao Yang, and Hongyu Zhao.

11:30 – 12: 00 Qiongshi Lu, Biostatistics, Yale University

Title: Post-GWAS prioritization through integrated analysis of genomic functional annotation

Abstract: Genome-wide association study (GWAS) has been a great success in the past decade, with tens of thousands of loci identified associated with many complex diseases in humans. However, significant challenges still remain in both identifying new risk loci and interpreting results. Bonferroni-corrected significance level is known to be conservative, leading to insufficient statistical power when the effect size is small to moderate at risk locus. Complex structure of linkage disequilibrium also makes it challenging to separate causal variants from nonfunctional ones in large haplotype blocks. We describe GenoWAP (Genome Wide Association Prioritizer), a post-GWAS prioritization method that integrates genomic functional annotation and GWAS test statistics. The effectiveness of GenoWAP is demonstrated through its applications to GWAS results for Crohn's disease and schizophrenia using the largest studies available. After prioritization based on a subset of all the available samples, highly ranked loci show substantially stronger signals in the whole dataset than the top loci before prioritization. At the single nucleotide polymorphism (SNP) level, top ranked SNPs after prioritization have both higher replication rates and consistently stronger enrichment of eQTLs. Within each risk locus, GenoWAP is able to distinguish real signal sources from groups of correlated SNPs. The GenoWAP software is available at http://genocanyon.med.yale.edu/GenoWAP

12: 00 – 14:00 Lunch break

14:00 – 14:30 Shuangge Ma, Biostatistics, Yale University

Title: Promoting Similarity of Sparsity Structures in Integrative Analysis

Abstract: For data with high-dimensional covariates but small to moderate sample sizes, the analysis of single datasets often generates unsatisfactory results. The integrative analysis of multiple independent datasets provides an effective way of pooling information and outperforms single-dataset analysis and some alternative multi-datasets approaches including meta-analysis. Under certain scenarios, multiple datasets are expected to share common important covariates, that is, their models have similarity in sparsity structures. However, the existing methods do not have a mechanism to promote the similarity of sparsity structures in integrative analysis. In this study, we consider penalized variable selection and estimation in integrative analysis. We develop a penalization based approach, which is the first to explicitly promote the similarity of sparsity structures. Computationally it is realized using a coordinate descent algorithm. Theoretically it has the much desired consistency properties. In simulation, it significantly outperforms the competing alternative when the models in multiple datasets share common important covariates. It has better or similar performance as the alternative when there is no shared important covariate. Thus it provides a “safe” choice for data analysis. Applying the proposed method to three lung cancer datasets with gene expression measurements leads to models with significantly more similar sparsity structures and better prediction performance.

14: 30 – 15:00 Bin Nan, Statistics, University of Michigan

Title: Large covariance/correlation matrix estimation for temporal data

Abstract: We consider the estimation of high-dimensional covariance and correlation matrices under slow-decaying temporal dependence. For generalized thresholding estimators, convergence rates are obtained and properties of sparsistency and sign-consistency are established. The impact of temporal dependence on convergence rates is also investigated. An intuitive cross-validation method is proposed for the thresholding parameter selection, which shows good performance in simulations. Convergence rates are also obtained for banding method if the covariance or correlation matrix is bandable. The considered temporal dependence has longer memory than those in the current literature and has particular implications in analyzing resting-state fMRI data for brain connectivity studies. This is a joint work with Hai Shu.

15:00 – 15:20 Break

15:20 – 15: 50 Ruoqing Zhu

Title: Greedy Tree Learning of Optimal Personalized Treatment Rules

Abstract: We propose a subgroup identification approach for detecting optimal personalized treatment rules with high-dimensional covariates. We adopt a greedy tree algorithm to pursuit signals in a high-dimensional space and yield interpretable optimal treatment rules. The proposed method consists of two steps. In the first step, we transform the subgroup identification problem into a weighted Classification problem that can utilize tree-based methods. In the second step, we adopt a newly proposed method, reinforcement learning trees, to detect features involved in the optimal treatment rules and construct binary splitting rules. The method is also extended to right censored survival data by using the accelerated failure time model and introducing double weighting to the classification trees. The performance of the proposed method is demonstrated via simulation studies and analyses of the Cancer Cell Line Encyclopedia (CCLE) data.

15: 50 – 16: 20 Tiejun Tong

Title: Bias and variance reduction in estimating the proportion of true null hypotheses

Abstract: When testing a large number of hypotheses, estimating the proportion of true nulls, denoted by pi0, becomes increasingly important. This quantity has many applications in practice. For instance, a reliable estimate of pi0 can eliminate the conservative bias of the Benjamini–Hochberg procedure on controlling the false discovery rate. It is known that most methods in the literature for estimating pi0 are conservative. Recently, some attempts have been paid to reduce such estimation bias. Nevertheless, they are either over bias corrected or suffering from an unacceptably large estimation variance. In this paper, we propose a new method for estimating pi0 that aims to reduce the bias and variance of the estimation simultaneously. To achieve this, we first utilize the probability density functions of false-null p-values and then propose a novel algorithm to estimate the quantity of pi0. The statistical behavior of the proposed estimator is also investigated. Finally, we carry out extensive simulation studies and several real data analysis to evaluate the performance of the proposed estimator. Both simulated and real data demonstrate that the proposed method may improve the existing literature significantly.

July 7

Room 1217, Department of Mathematics, Hong Kong Baptist University

9:30- 10:00 Hongyu Zhao, Biostatistics, Yale University

Title: Spatial Temporal Modeling of Gene Expression Dynamics During Human Brain Development

Abstract: Human neurodevelopment is a highly regulated biological process. Recent technological advances allow scientists to study the dynamic changes of neurodevelopment at the molecular level through the analysis of gene expression data from human brains. In this talk, we focus on the analysis of data sampled from 16 brain regions in 15 time periods of neurodevelopment. We will introduce a two-step statistical inferential procedure to identify expressed and unexpressed genes and to detect differentially expressed genes between adjacent time periods. Markov Random Field (MRF) models are used to efficiently utilize the information embedded in brain region similarity and temporal dependency in our approach. We will also describe a Bayesian neighborhood selection procedure to estimate Gaussian Graphical Models (GGMs) across time and space. We have developed and implemented an efficient algorithm for statistical inference. Simulation studies suggest that our approach achieves better accuracy in network estimation compared with models not incorporating spatial and temporal dependency. We also show the graph selection consistency of the proposed method in the sense that the posterior probability of the true model converges to one. This is joint work with Zhixiang Lin, Tao Wang, Can Yang, Stephan Sanders, Mingfeng Li, Nenad Sestan, and Matthew State.

10:00 – 10:30 Ji Zhu, Statistics, University of Michigan

Title: Detecting Overlapping Communities in Networks with Spectral Methods

Abstract: Community detection is a fundamental problem in network analysis. In practice, it often occurs that the communities overlap, which makes the problem more challenging. Here we propose a general, flexible, and interpretable generative model for overlapping communities, which can be thought of as a generalization of the degree-corrected stochastic block model. We develop an efficient spectral algorithm for estimating the community memberships, which deals with the overlaps by employing the K-medians algorithm rather than the usual K-means for clustering in the spectral domain. We show that the algorithm is asymptotically consistent when networks are not too sparse and the overlaps between communities not too large. Numerical experiments on both simulated networks and many real social networks demonstrate that our method performs well compared to a number of benchmark methods for overlapping community detection. This is joint work with Yuan Zhang and Elizaveta Levina.

10:30 – 10:40 Break

10:40 – 11: 40 Discussion