The Chinese University of Hong Kong
Department of Statistics

Seminar

Variable Selection for High-dimensional Data Using Known and Novel Graph Information

By

Professor Qi LONG
Department of Biostatistics, Epidemiology and Informatics
University of Pennsylvania

Abstract

Variable selection for structured high-dimensional covariates lying on an underlying graph (say, representing biological pathways) has drawn considerable interest. However, most of the existing methods may not be scalable to high dimensional settings involving tens of thousands of variables lying on known pathways such as the case in genomics studies, and they assume that the graph information is fully known. This talk will focus on addressing these two challenges. In the first part, I will present an adaptive Bayesian shrinkage approach which incorporates known graph information through shrinkage parameters and is scalable to high dimensional settings (e.g., $p \sim 100{,}000$ or millions). We also establish theoretical properties of the proposed approach for fixed and diverging p. In the second part, I will tackle the issue that graph information is not fully known. For example, the role of miRNAs in regulating gene expression is not well-understood and the miRNA regulatory network is often not validated. We propose an approach that treats unknown graph information as missing data (i.e. missing edges), introduce the idea of imputing the unknown graph information, and define the imputed information as the novel graph information. In addition, we propose a hierarchical group penalty to encourage sparsity at both the pathway level and the within-pathway level, which, combined with the imputation step, allows for incorporation of known and novel graph information. The methods are assessed via simulation studies and are applied to analyses of cancer data.

Date:　　March 7, 2018 (Wednesday)
Time:　　2:30 p.m. - 3:30 p.m.
Venue:　　Lee Shau Kee Building (LSK) - Room LT2
　　　　　The Chinese University of Hong Kong

ALL INTERESTED ARE WELCOME