Variable Selection for Multiple Types of High-Dimensional Features With Missing Data

Kin Yau Wong

The Hong Kong Polytechnic University
E-mail: kin-yau.wong@polyu.edu.hk

Abstract: Recent technological advances have made it possible to collect multiple types of high-dimensional data in biological, clinical, and epidemiological studies. However, some data types or features may not be measured for all study subjects because of cost or other constraints. A common strategy for handling incomplete (high-dimensional) data is to obtain a complete data set using listwise deletion or through single imputation and then apply conventional statistical methods to the complete data set. This two-step approach, however, is inefficient and may even be biased. In this presentation, we present a valid and efficient approach to variable selection with multiple types of potentially missing features. We use a latent variable model to characterize the relationships across and within data types and to infer missing values from observed data. We develop a penalized-likelihood approach for variable selection and parameter estimation and devise an efficient expectation-maximization (EM) algorithm to implement our approach. The likelihood-based framework accommodates general missing-data patterns, and the low-dimensional factor model makes the estimation computationally tractable. We provide an application to a motivating multi-platform genomics study.