“Variable selection for high-dimensional latent class models”

PI: Guan-Hua Huang (National Chiao Tung University)

Role: Principal investigator

Ministry of Science and Technology of Taiwan: 08/01/2018-07/31/2020

Latent class models (LCMs) are widely used in analyzing multiple measured variables. In modern genomic research (e.g., high-throughput microarray and SNP array studies), a large number of variables are measured from moderate or small samples. Analysis of these high-dimensional data is difficult due to many noisy and overlapping features in such data, which can disturb the results. Thus, it is important to carry out some form of the variable selection made before or incorporated into the fitting procedure to exclude noisy and uninformative variables. In this project, we define variables that have no difference among unobservable latent classes as “noisy” variables, and other variables that have different distributions in different latent classes as “clustering” variables. We propose an “alternate k-means method” to identify these noisy variables and exclude their influences in estimating latent classes. We will evaluate the performance of the proposed method and compare it with competing approaches via simulations. Microarray gene expression data, the traffic flow data on the Freeway No. 5 in Taiwan, and schizophrenia positive and negative syndrome scale data will be analyzed to demonstrate the usefulness of the proposed method.

Keywords: high-dimensional data; k-means clustering; latent class model; microarray gene expression; traffic flow prediction; variable selection.