Bayesian Variable Selection for Finite Mixture Model of Linear Regressions

Ray-Bing Chen
Department of Statistics, National Cheng Kung University

Abstract

We propose a Bayesian method for variable selection in the finite mixture model of linear regressions. The model assumes that the observations come from a heterogeneous population which is a mixture of a finite number of sub-populations. Within each sub-population, the response variable can be explained by a linear regression on the predictor variables. So the whole data set can be modeled by a mixture of linear regressions, where each mixture component follows a separate regression model. In the case where the number of predictor variables is large, it is assumed that only a small subset of variables are important for explaining the response variable. It is further assumed that for different mixture components, different subsets of variables may be needed to explain the response variable. This gives rise to a complex variable selection problem. We propose to solve this problem within the Bayesian framework where we introduce two sets of latent variables. For the first set of latent variables, each observation is associated with an indicator, indicating which sub-population or mixture component this observation comes from. For the second set of latent variables, within each mixture component, each predicator variable is associated with an indicator, indicating whether this variable is included in the regression model of the mixture component. Variable selection can then be accomplished by sampling from the posterior distribution of the indicators as well as the coefficients of the selected variables. We conduct simulation studies to demonstrate that the proposed method performs well in comparison with existing methods. We also analyze two real data sets to further illustrate the proposed method.