Identifying Cluster Structures and Relevant Variables in High-Dimensional Datasets

Mahlet G. Tadesse
Department of Mathematics and Statistics,
Georgetown University, Washington, DC

Abstract:

In analyzing high-dimensional datasets, there is often interest in uncovering cluster structures and identifying variables associated with the clusters. I will present some Bayesian methods we have proposed to address such questions in a unified manner. The first problem I will discuss is concerned with discovering homogeneous subgroups of samples and identifying variables that discriminate across the subgroups. We use mixture models with an unknown number of components to uncover the cluster structures and build a stochastic search variable selection method into the model to identify discriminating variables. The second problem is concerned with relating two high-dimensional data sets by uncovering cluster structures in the data and identifying groups of associated variables across the data sets. We use a stochastic partitioning method that combines ideas of mixtures of regression models and variable selection methods to search for sets of covariates associated with sets of correlated outcomes. I will illustrate the methods with applications to genomic data sets.

Keywords: Bayesian inference; genomic studies; high-dimensional data; Markov chain Monte Carlo; mixture models; variable selection.