High-Dimensional Clustering via a Latent Transformation Mixture Model

Hui Zou

School of Statistics, University of Minnesota

ABSTRACT

Cluster analysis is a fundamental task in machine learning. From the probabilistic modeling viewpoint, a finite mixture model is naturally suited for the distribution of data with multiple clusters, and hence model-based clustering (MBC) offers an effective solution. Despite its many successful applications, MBC also often underperforms due to its potential severe modeling bias. We aim to design a more robust off-the-shelf MBC for high-dimensional data by mitigating the model bias. To this end, we propose a novel CESME model by incorporating nonparametric latent transformations into the finite Gaussian mixture model (GMM). The inclusion of latent transformations significantly enhances the flexibility of the finite GMM without compromising interpretability. We derive a model fitting procedure for implementing the optimal clustering under the CESME model and analyze the clustering accuracy of the resulting algorithm. It is shown that the additional cost due to estimating nonparametric transformations is negligible compared with an ideal clustering algorithm with known transformations. On six benchmark single-cell RNA sequence datasets, CESME exhibits dominating advantages over existing methods in the literature.

Keywords: Model-based clustering, Nonparametric transformation, High-dimensional data