Infinite mixtures of infinite factor analysers: a model-based approach to clustering high dimensional data.

Keefe Murphy[1], Cinzia Viroli[2], and Isobel Claire Gormley[1]

[1]University College Dublin
[2]Università di Bologna

Abstract

Gaussian mixture models with a factor-analytic structure are often employed as a model-based approach to clustering high-dimensional data. Typically, the numbers of clusters and latent factors must be specified in advance of model fitting, and the optimal pair selected using a model choice criterion. For computational reasons, models in which the number of latent factors is common across clusters are generally considered.

Here the infinite mixture of infinite factor analysers (IMIFA) model is introduced. IMIFA employs a Poisson-Dirichlet process prior to facilitate automatic inference on the number of clusters. Further, IMIFA employs shrinkage priors to allow cluster specific numbers of factors, automatically inferred via an adaptive Gibbs sampler. IMIFA is presented as the flagship of a family of factor-analytic mixture models, providing flexible approaches to clustering high-dimensional data.

Applications to benchmark and real data sets illustrate the IMIFA model and its advantageous features: IMIFA obviates the need for model selection criteria, reduces model search and associated computational burden, improves clustering performance by allowing cluster-specific numbers of factors, and quantifies uncertainty in the numbers of clusters and cluster-specific factors.

Implementation of the proposed methodology is facilitated through the associated open source R package IMIFA.