Inferring components and clusters in Bayesian finite mixture modeling

Bettina Grün

Abtract:

Bayesian cluster analysis aims at inferring the number of data clusters present in a data set. This can be achieved using either finite or infinite mixture models. In Bayesian nonparametrics the number of data clusters is determined using the realized partitions. In Bayesian finite mixture models usually a one-to-one relationship between components and data clusters is assumed. The posterior of the number of components is then approximated using different methods, e.g., reversible jump Markov chain Monte Carlo (Richardson and Green, 1997), Markov birth-and-death process sampling (Stephens, 2000) or the Jain-Neal split-merge sampler as proposed by Miller and Harrison (2018).

In the framework of finite mixture models we propose to explicitly distinguish between the number of data clusters and the number of components and purposely allow for more components than data clusters. We extend the standard Bayesian finite mixture model by including priors on the number of components and on the Dirichlet parameter. This allows us to approximate the posteriors of the number of components as well as data clusters using Gibbs sampling techniques. The performance of the proposed sampling technique is compared to previously proposed approaches. The additional flexibility gained by suitably selecting the parameters of the hyperpriors is highlighted and guidance for their choice provided.