Bayesian Mixtures for Cluster Analysis and Flexible Modeling of Distributions

Dissertation by Arno Fritsch

Abstract: Finite mixture models assume that a distribution is a combination of several parametric distributions. They offer a compromise between the interpretability of parametric models and the flexibility of nonparametric models. This thesis considers a Bayesian approach to these models, which has several advantages. For example, using only weak prior information, it can solve problems with unbounded likelihood functions, that can occur in mixture models. The Bayesian approach also allows an elegant extension of finite to (countable) infinite mixture models. Depending on the application, the components of mixture models can either be viewed as just a means to the flexible modeling of a distribution or as defining subgroups of a population with different parametric distributions. Regarding the former case consistency results for Bayesian mixtures are stated. An example concerning the flexible modeling of a random effects distribution in a logistic regression is also given. The application considers the goalkeeper's effect in saving a penalty. In the latter case mixture models can be used for clustering. Bayesian mixtures then allow the estimation of the number of clusters at the same time as the cluster-specific parameters. For cluster analysis the standard approach for fitting Bayesian mixtures, Markov Chain Monte Carlo (MCMC), unfortunately leads to inferential difficulties. The labels associated with the clusters can change during the MCMC run, a phenomenon called label-switching. The problem gets severe, if the number of clusters is allowed to vary. Existing methods to deal with label-switching

and a varying number of components are reviewed and new approaches are proposed for both situations. The first consists of a variant of the relabeling algorithm of Stephens (2000). The variant is more general, as it applies to drawn clusterings and not drawn parameter values. Therefore it does not depend on the specific form of the component distributions. The second approach is based on pairwise posterior probabilities and is an improvement of a commonly used loss function due to Binder (1978). Minimization of this loss is shown to be equivalent to maximizing the posterior expected Rand index with the true clustering. As the adjusted Rand index is preferable to the raw index, the maximization of the posterior expected adjusted Rand is proposed. The new approaches are compared to the previous methods on simulated and real data. The real data used for cluster analysis are two gene expression data sets and Fisher's iris data.

References:

- Binder, D. A. (1978). "Bayesian Cluster Analysis." *Biometrika*, 65: 31–38.

- Stephens, M. (2000). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society, Series B*, 62: 795–809.