Model-based clustering based on sparse finite Gaussian mixtures

Gertraud Malsiner-Walli[1], **Sylvia Frühwirth-Schnatter**[2], **Bettina Grün**[1]

1. Department of Applied Statistics, Johannes Kepler University Linz, Austria
2. Institute for Statistics and Mathematics, Vienna University of Economics and Business, Austria

In model-based clustering selecting a suitable number of components for a finite mixture distribution is a challenging problem. We want to contribute to this issue and propose a new Bayesian approach. We propose to deliberately overfit the mixture model by specifying K^{max} components, K^{max} being a reasonable upper bound on the number of components. Simultaneously, we specify a sparse hierarchical prior on the component weights which has the effect of emptying superfluous components during MCMC sampling. A straightforward criterion for estimating the true number of components is given by the most frequent number of nonempty groups visited during MCMC sampling. In addition, we also examine the effect of specifying a sparse hierarchical prior on the component means, namely the normal gamma prior. By this we aim at reducing the MSE of estimated parameters in the presence of noise variables. We perform Bayesian estimation of finite Gaussian mixture models using MCMC methods based on data augmentation and Gibbs sampling. To obtain an identified mixture model, in a post-processing step the MCMC output is relabeled using k-centroid cluster analysis based on the Mahalanobis distance. We evaluate our proposed method in a simulation setup with artificial data and by applying it to benchmark data sets.