A toolbox for Bayesian model-based clustering in R

Sylvia Frühwirth-Schnatter[1], Jan Greve[1], Bettina Grün[2] and Gertraud Malsiner-Walli[1]

[1]WU Vienna University of Economics and Business; [2]Johannes Kepler University Linz

Implementation of comprehensive computational tools for a specific statistical methodology is often as important as its development for their uptake by users. In this matter, the R environment for statistical computing and graphics (R Core Team, 2019) has become the lingua franca for statistical computing but also for quantitative research methods in general. Specifically, for model-based clustering, several packages are available from the Comprehensive R Archive Network (CRAN) which provide the necessary tools for model fitting as well as inspection and visualization in a frequentist setting. These include, among others, package **mclust** (Scrucca et al. 2016) for finite mixtures of multivariate Gaussian distributions and package **flexmix** (Grün and Leisch 2008) for finite mixtures of regression models.

However, for Bayesian model-based clustering, the set of available implementations in R is less comprehensive and scattered. Frühwirth-Schnatter (2006) provides a thorough introduction into the theory and the application of Bayesian mixture modeling, accompagnied by supplementary material consisting of a software implementation. Nonetheless, this software toolbox is implemented in MATLAB. We aim at providing an R package which facilitates and streamlines Bayesian model-based clustering and also incorporates recent advances in Bayesian mixture modeling including the use of shrinkage priors to obtain sparse solutions.

The package guides users through all necessary steps such as model specification, prior choice, sampling, post-processing and inference. The user-friendly object-oriented implementation facilitates suitable solutions with minimal required user-input with the additional capacity to inspect and explore obtained solutions. At the same time, expert-specific features such as the possibility to exchange default prior specifications to another set of appropriate priors are also supported. Finally, a modular implementation will be pursued which allows to easily extend the toolbox to cover "new" mixture models including different component distributions, revert to different sampling engines to obtain the MCMC draws and explore the use of different post-processing tools for model identification.

References

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, New York. DOI:10.1007/978-0-387-35768-3.

Grün, B. and Leisch, F. (2008) FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. *Journal of Statistical Software*, 28(4), 1–35. DOI:10.18637/jss.v028.i04.

R Core Team (2019). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL:https://www.R-project.org/.

Scrucca, L., Fop, M., Murphy, T. B. and Raftery, A. E. (2016) **mclust** 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1), 205–233. DOI:10.32614/RJ-2016-021.