Clustering-based models in R, with application on univariate Gaussian mixtures: review, evaluation and extensions

Bastien CHASSAGNOL[*] Etienne BECHT[†] Mickaël GUEDJ[‡]
Pierre-Henri WUILLEMIN[§] Gregory NUEL[¶]

Mots-clefs : modèles de mélange – clustering – algorithme EM – benchmark

Finite mixture models are increasingly used for modeling and dealing with stochastic problems such as clustering, classification and regression, with many applications in biological fields. Unsurprisingly, many packages have been developed to fit mixture models, arising the natural question which is the best suited one, depending on the use case.

However, to our knowledge, no review describing the main features offered by these packages and comparing their computational and statistical performances has been performed. In this talk, we focus on packages implementing the EM algorithm on univariate Gaussian mixture distributions, being the most common use case.

A mixture model aims at representing a distribution that can be split into a weighted sum of components, each of which representing an independent sub-population with specific characteristics. Besides, these sub-distributions are generally unknown. Common applications of mixture models include Chemometrics, DNA sequence analysis, Transcriptomics and Epidemiology. Mixture models can be used for clustering, and yields a parametric distribution for each cluster. This can be achieved using the EM algorithm, which has been shown to be consistent in (Dempster, Laird, and Rubin 1977).

A systematic research on CRAN has led us to compare eight packages that directly deal with clustering data with underlying assumption of Gaussian distribution for each component: *flexmix*(Grün and Leisch 2008), *bgmm*(Biecek et al. 2012), *Emcluster*(Chen et al. 2019), *mclust*(Scrucca et al. 2016), *mixtools*(Benaglia et al. 2009), *Rmixmod*(Lebret et al. 2015), *HDClassif*(Bergé, Bouveyron, and Girard 2012) and *mixture*(Pocuca, Browne, and McNicholas 2021). We describe the main features offered by these packages, focusing on comparing these packages for evaluating Gaussian mixture distributions. Additionally, as a baseline for our benchmark results, we re-implemented the EM algorithm in the base R language. We highlight differences in these packages' capabilities and design choices. For each package, we reviewed and benchmarked the algorithms estimating the parameters of the mixture model, the flexibility of parametrization of the component distributions, the flexibility and default strategy in the initialization of the algorithm, and the specific heuristics of each package compared to the default EM algorithm. These heuristics aim at avoiding solutions corresponding to local minima, and adding a prior information on the distribution of the parameters.

Using simulations, we extensively compared the performance of these packages in terms of bias and variance of their estimations and in their running times. We evaluated the impact of the number of components, the degree of imbalance, the presence of outliers, the skewness of the mixtures, and the level of discrimination between components.

We also comprehensively evaluated the performance of each initialization method with the optimization procedure of each package. We show that the choice of initialization method has a strong impact on performance.

[*] LPSM, UMR CNRS 8001, 4 Place Jussieu Sorbonne University, Paris, bastien.chassagnol@upmc.fr
[†] Servier, 50 Rue Carnot, 92150, Suresnes, etienne.becht@servier.com
[‡] Servier, 50 Rue Carnot, 92150, Suresnes, mickael.guedj@servier.com
[§] LIP6, UMR7606, 4 Place Jussieu Sorbonne University, Paris, pierre-henri.wuillemin@lip6.fr
[¶] LPSM, UMR CNRS 8001, 4 Place Jussieu Sorbonne University, Paris, Gregory.Nuel@math.cnrs.fr

The quantiles or kmeans methods gave the best results with balanced and well-separated components. Conversely, random initialization performs better when these assumptions are not met. Interestingly, while all these packages implement the same deterministic EM algorithm, comparison to our base R implementation highlighted package-specific implementation details that made them sensitive to the choice of initialization method. More precisely, the *bgmm*, *Emcluster* and *mclust* implementations are more sensitive to the quality of the initialization, while *mixtools*, *Rmixmod* and base R implementation of the original EM algorithm are more flexible but with a higher variance on the estimated of the parameters.

This work highlights the differences among available R packages implementing the EM algorithm and suggest the best use cases for each package as well as their limitations.

Références

Benaglia, Tatiana, Didier Chauveau, David R. Hunter, and Derek S. Young. 2009. "Mixtools: An R Package for Analyzing Finite Mixture Models." *Journal of Statistical Software* 32 (6): 1–29. https://hal.archives-ouvertes.fr/hal-00384896.

Bergé, Laurent, Charles Bouveyron, and Stéphane Girard. 2012. "HDclassif: An R Package for Model-Based Clustering and Discriminant Analysis of High-Dimensional Data." *Journal of Statistical Software* 46 (1, 1): 1–29. https://doi.org/10.18637/jss.v046.i06.

Biecek, Przemyslaw, Ewa Szczurek, Martin Vingron, and Jerzy Tiuryn. 2012. "The R Package Bgmm: Mixture Modeling with Uncertain Knowledge." *Journal of Statistical Software* 47 (1, 1): 1–31. https://doi.org/10.18637/jss.v047.i03.

Chen, Wei-Chen, Ranjan Maitra, Volodymyr Melnykov, Dan Nettleton, David Faden, Rouben Rostamian, and R. Core team (some functions are modified from the R. source code). 2019. *EMCluster: EM Algorithm for Model-Based Clustering of Finite Mixture Gaussian Distribution* (version 0.2-12). https://CRAN.R-project.org/package=EMCluster.

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39 (1): 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x.

Grün, Bettina, and Friedrich Leisch. 2008. "FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters." *Journal of Statistical Software* 28 (1, 1): 1–35. https://doi.org/10.18637/jss.v028.i04.

Lebret, Rémi, Serge Iovleff, Florent Langrognet, Christophe Biernacki, Gilles Celeux, and Gérard Govaert. 2015. "Rmixmod: The R Package of the Model-Based Unsupervised, Supervised, and Semi-Supervised Classification Mixmod Library." *Journal of Statistical Software* 67 (1, 1): 1–29. https://doi.org/10.18637/jss.v067.i06.

Pocuca, Nik, Ryan P. Browne, and Paul D. McNicholas. 2021. *Mixture: Mixture Models for Clustering and Classification* (version 2.0.3). https://CRAN.R-project.org/package=mixture.

Scrucca, Luca, Michael Fop, T. Brendan Murphy, and Adrian E. Raftery. 2016. "Mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models." *The R Journal* 8 (1): 289–317. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5096736/.