New mixture models and algorithms in the mixtools package

Didier Chauveau

New mixture models and algorithms in the mixtools package

Didier Chauveau

MAPMO - Fédération Denis Poisson
Université d'Orléans and CNRS UMR 7349
BP 6759, 45067 Orléans cedex 2
didier.chauveau@univ-orleans.fr

Mots clefs : Finite mixture, nonparametric mixtures, EM algorithms

The mixtools package for the R statistical software [7] has evolved from 2006 up to the current CRAN version. Benaglia et al. [2] give a comprehensive account of mixtools capabilities as in 2009. This package provide various tools for analyzing a variety of finite mixture models, from traditional methods such as EM algorithms for uni- and multivariate Gaussian mixtures, up to more specific and recent models such as, e.g., multinomial mixtures, mixtures of regression or multivariate non-parametric mixtures.
Since then, new and different models connected to mixtures have been investigated by several authors, some of those involved in mixtools' development. For most of these model analysis, new or specific computational techniques have been progressively implemented in the development version of the package, taking advantage of its environnement. This talk, that involves joint works with the co-authors cited below, presents some of these models and illustrate mixtools' new capabilities that have been added since the publication of Benaglia et al. [2]. These new models share in common the description of the distribution of the observations by a finite mixture density

$$g(x|\boldsymbol{\theta}) = \sum_{j=1}^{m} \lambda_j f_j(x), \quad \boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{f}), \quad x \in \mathbb{R}^r,$$

where $\boldsymbol{\theta}$ is the model parameter, consisting in the *component densities* f_j's and component weights λ_j's that are positive and sum to unity. Precise specification of $\boldsymbol{f}$ depends on the model assumptions, e.g., for univariate normal mixtures f_j is the density of $\mathcal{N}(\mu_j, \sigma_j^2)$ and $\boldsymbol{f} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$, the m-vectors of component means and variances.

Gaussian mixtures with constrained parameters: Motivated by mixture models issued from psychometrics, Chauveau and Hunter [5] consider the problem of linear constraints on the parameters $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ for finite mixtures of normal components. Surprisingly, we show that even for simple linear constraints on $\boldsymbol{\mu}$ such as $\boldsymbol{\mu} = M\boldsymbol{\beta}+C$ for some unknown p-vector $\boldsymbol{\beta}$ with $p \leq m$, and known matrix M and vector $\boldsymbol{C}$, the Maximum Likelihood Estimation problem succumbs to an ECM (with Conditional-M steps) generalization of the EM algorithm. With certain types of variance constraints, a further generalization of EM known as MM (Majorization-Minorization) algorithm have also been added in mixtools.

Nonparametric MM for smoothed likelihood maximization: Benaglia et al. [1] originally designed an empirical "nonparametric EM" (npEM) algorithm for fitting multivariate nonparametric mixtures with completely unspecified component densities except for a conditional independence assumption $f_j(x) = \prod_{k=1}^{r} f_{jk}(x_k)$ which means that the (scalar) coordinates of the r-dimensional observation x are independent conditional on the component from which x

is drawn. Despite its superiority over competing methods shown by numerical evidence, this npEM algorithm which is in the spirit of an EM in its formulation lacks any sort of theoretical justification. Following this work, Levine et al. [6] have proposed and implemented a new MM algorithm which does provide an ascent property (just as a genuine EM does) with respect to a smoothed loglikelihood, at the cost of a higher computing load. Both versions are know available in mixtools.

Reliability mixture models on randomly censored data: Mixtures are also suitable to modelize lifetime data, but these data are often censored. Randomly censored data from mixture models have been considered in Bordes and Chauveau [3], both for parametric or semiparametric mixtures. They propose several algorithms, from genuine parametric EM for specific families, to parametric and semiparametric Stochastic EM (St-EM). These stochastic versions, that include an additional step for simulating the missing part of the data, provide workable estimation methods since completion of the data for the component indicators allows application of nonparametric estimates for survival data such as the Kaplan-Meier estimate. Most of these algorithms are already implemented in the development version of mixtools.

Semiparametric EM with one component known: In multiple testing and False Discovery Rate estimation, semiparametric mixtures with one component known can be used (Bordes et al. [4]). In Saby et al. [8], some new EM-like algorithms of this kind have been implemented in mixtools and tested on simulated and actual data.

References

[1] Benaglia, T., Chauveau, D., and Hunter, D. R. (2009a). An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics*, 18(2):505–526.

[2] Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. (2009b). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29.

[3] Bordes, L. and Chauveau, D. (2010). Some algorithms to fit some reliability mixture models under censoring. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of the 19th International Conference on Computational Statistics, Paris France*. Springer.

[4] Bordes, L., Delmas, C., and Vandekerkhove, P. (2006). Semiparametric estimation of a two-component mixture model where one component is known. *Scand. J. Statistics*, 33:733–752.

[5] Chauveau, D. and Hunter, D. R. (2011). ECM and MM algorithm for mixtures with constrained parameters. Technical Report hal-00625285, version 1, HAL.

[6] Levine, M., Hunter, D. R., and Chauveau, D. (2011). Smoothed likelihood for multivariate mixtures. *Biometrika*, 98(2):403–416.

[7] R Development Core Team (2010). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

[8] Saby, N., Orton, T. G., Chauveau, D., Lemercier, B., Walter, C., Schvartz, C., and Arrouays, D. (2011). Application of a mixture model approach to large-scale simultaneous hypothesis testing in soil monitoring. In *Pedometrics*.