ECM and MM algorithms for normal mixtures with constrained parameters

Didier Chauveau, David R. Hunter

ECM and MM algorithms for normal mixtures with constrained parameters

Didier Chauveau[1] David R. Hunter[2]

[1]MAPMO - CNRS UMR 7349, Université d'Orléans, France
[2]Pennsylvania State University, PA, USA

August 7, 2013

Abstract: EM algorithms for obtaining maximum likelihood estimates of parameters in finite normal mixture models are well-known, and certain types of constraints on the parameter space, such as the equality of variance assumption, are very common. Here, we consider more general constraints on the parameter space for finite mixtures of normal components. Surprisingly, these simple extensions have not been explored in the literature. We show how the MLE problem yields to an EM generalization known as an ECM algorithm. For certain types of variance constraints, yet another generalization of EM, known as MM algorithms, is required. After a brief explanation of these algorithmic ideas, we demonstrate how they may be applied to parameter estimation and hypothesis testing in finite mixtures of normal components in the presence of linear constraints on both mean and variance parameters. We provide implementations of these algorithms in the mixtools package for the R statistical software.

Keywords: generalized EM algorithms, ECM algorithms, MM algorithms, finite mixture, Likelihood ratio tests.

1 Introduction

Finite mixture models give a flexible way to model a wide variety of random observations (see, e.g., McLachlan and Peel, 2000). In such models, we assume that n independent measurements $X_1, \ldots, X_n$ are observed such

that each X_i comes from one of m possible component distributions. Importantly, the component number, 1 though m, is not observed along with X_i. Notationally, it is common to define the (unobserved) indicator variables

$$Z_{ij} = I\{\text{observation } i \text{ comes from component } j\},$$

where it is assumed that, unconditional on X_i, each Z_{ij} has expectation λ_j.

Throughout this article, we shall assume that each component distribution has a density with respect to Lebesgue measure that is known up to the value of a parameter. Thus, the density of each X_i may be written

$$g_{\boldsymbol{\theta}}(x) = \sum_{j=1}^{m} \lambda_j f(x; \xi_j), \tag{1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\xi}) = (\lambda_1, \ldots, \lambda_m, \boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m)$ denotes the parameter, and the λ_j are positive and sum to unity. (We disallow the possibility that any $\lambda_j = 0$ in this context.) For simplicity this article will focus on the univariate normal case for which $f(x; \boldsymbol{\xi}_j)$ is the normal $\mathcal{N}(\mu_j, \sigma_j^2)$ density.

The EM algorithm, as defined in the seminal paper of Dempster et al. (1977), is more properly understood to be a class of algorithms, a number of which predate even the Dempster et al. (1977) paper in the literature. These algorithms are designed for maximum likelihood estimation in missing data problems, and finite mixture problems are canonical examples of these problems because the unobserved Z_{ij} give an easy interpretation of missing data. For a comprehensive and recent account of EM algorithms, refer to McLachlan and Krishnan (2008).

If we consider $(X_1, Z_1), \ldots, (X_n, Z_n)$ to be the complete data in a finite mixture example where only the X_i are actually observed, the corresponding EM algorithm consists of writing the complete data log-likelihood function

$$L_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{j=1}^{m} Z_{ij} \log\left(\lambda_j f(x_i; \boldsymbol{\xi}_j)\right), \tag{2}$$

as well as its expectation under the assumption that the parameter governing the random behavior of Z_{ij} at iteration t is $\theta^{(t)}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\boldsymbol{\theta}^{(t)}}\left[L_c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{Z}) \mid \mathbf{X} = \mathbf{x}\right] = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(t)} \log\left(\lambda_j f(x_i; \boldsymbol{\xi}_j)\right), \tag{3}$$

where

$$p_{ij}^{(t)} = \mathbb{P}_{\boldsymbol{\theta}^{(t)}}\left(Z_{ij} = 1 \mid X_i = x_i\right) = \frac{\lambda_j^{(t)} f(x_i; \boldsymbol{\xi}_j^{(t)})}{\sum_{r=1}^{m} \lambda_r^{(t)} f(x_i; \boldsymbol{\xi}_r^{(t)})} \tag{4}$$

is often called the posterior probability of individual i coming from component j, given the current $\boldsymbol{\theta}^{(t)}$ and the observed x_i.

The iteration $\boldsymbol{\theta}^{(t)} \to \boldsymbol{\theta}^{(t+1)}$ is defined in this general setup by

1. E-step: compute $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ or, equivalently, compute $p_{ij}^{(t)}$ by (4).

2. M-step: set $\boldsymbol{\theta}^{(t+1)} = \text{argmax}_{\boldsymbol{\theta}}\, Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

Conveniently, the M-step for finite mixture models always looks partly the same: No matter what form f takes, the updates to the mixing proportions are given by

$$\lambda_j^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ij}^{(t)} \tag{5}$$

for $j = 1, \ldots, m$. The updates to the $\boldsymbol{\xi}$ parameters depend on the particular form of the component densities and will be discussed later.

The remainder of this article is organized as follows: Section 2 discusses the main idea of the paper, which is estimation in the presence of various constraints on the space of parameters in the finite mixture of normal distributions. Section 3 shows how an EM algorithm may be modified to handle a general class of constraints. In Section 4 we address the inferential question of the statistical test for a null hypothesis H_0 : "constraints hold" by a standard Likelihood Ratio Test (LRT) approach. Section 5 presents a simulation study in the spirit of a model from psychometrics and Section 6 summarizes the article.

2 Constraints on the parameter space

The case of normal component densities will be the sole topic of the remainder of this article, but extensions to any parametric family with mean and/or variance parameters should allow for similar ideas. In the present normal case, the density $f(x; \boldsymbol{\xi}_j)$ of Equation (1) is simply the normal density with parameters $\boldsymbol{\xi}_j = (\mu_j, \sigma_j^2)^\top$. As it happens, if we do not restrict the $\boldsymbol{\xi}$ parameters in any way, the likelihood resulting from a sample $x_1, \ldots, x_n$ is unbounded. This well-known problem (McLachlan and Peel, 2000) implies that no maximum likelihood estimator exists in the unconstrained problem. Various remedies are suggested in the literature, but perhaps the easiest, when it may be justified, is to impose the restriction that all σ_j are equal. It is straightforward to derive the EM algorithm for this equal-variances case.

Constrained EM in the literature To the best of our knowledge, multiple proportional constraints have not been handled specifically for parametric mixture models as we do here. In their recent book, McLachlan and Krishnan (2008, Section 3.5.4) give a brief account of what has been done so far. Constraints in mixture models have mostly been introduced to handle the difficulty of unboundedness of the likelihood function. Constraints such as $\sigma_i = c_{ij}\sigma_j$ in normal mixture models have been considered in Quandt and Ramsey (1978), but their approach was not connected to the EM algorithm. Hathaway (1985) and Hathaway (1986) reformulate the normal mixture problem into a constrained maximum-likelihood formulation (and EM algorithm) with similar constraints on the variances, in a way to exclude the points of degeneracy of the likelihood function. Nettleton (1999) studies convergence of the EM algorithm in parameter space with various inequality constraints, in a general framework. Kim and Taylor (1995) propose an EM algorithm under linear restrictions on the parameters, for general missing data models, using Newton-Raphson iterations within each M-step. In the same vein, Shi et al. (2005) consider linear constraints in a linear regression model with missing data. Equality and fixed-value constraints in the spirit of ours has been considered in Mooijaart and van der Heijden (1992), but for the case of categorical variables with latent class, using a Lagrange multiplier in the M-step.

In the multivariate normal case, there is a large literature on finite mixture models in which the covariance matrices satisfy various types of constraints. Following Banfield and Raftery (1993), several related papers such as Bensmail and Celeux (1996) and Fraley and Raftery (2002) use an eigenvalue decomposition of the component covariance matrices in which various pieces of the decomposition (e.g., the eigenvalues or eigenvector matrices) may be constrained in some way. In contrast, McNicholas et al. (2010) use a factor analysis decomposition, through which constraints may be imposed. In the univariate case that we consider, these decompositions do not apply since our covariances are not matrices but scalars.

2.1 Two definitions of constraints

Linear constraints for the mean vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^\top$ can be written

$$\boldsymbol{\mu} = M\boldsymbol{\beta} + \mathbf{C} \tag{6}$$

for some unknown p-vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ with $p \leq m$, and known $m \times p$ matrix M and fixed m-vector of constants $\mathbf{C}$. These sorts of constraints may appear to be straightforward to implement in a normal-mixture EM

algorithm. However, it turns out that these constraints are sometimes handled incorrectly. The most illustrative example of this phenomenon occurs in the model in which it is assumed that the means $\mu_1 = \cdots = \mu_m$ are equal, yet the variances are not all the same. In this case, the M-step does *not* consist of estimating μ_1 by the sample mean $\bar{x} = \sum_{i=1}^{n} x_i/n$ at each iteration, despite the fact that $\mu_1 = \mathbb{E}_{g_{\boldsymbol{\theta}}}(X)$. We explain in Section 3.1 why this is incorrect. Similar linear constraints for the variance parameters are discussed in Section 3.3.

Note that equation (6) is equivalent to a standard linear constraints of the form $A\boldsymbol{\mu} = b$, where A and b are known matrix and vector of adequate dimensions, but the expression of equation (6) with $\boldsymbol{\beta}$ as the parameter of interest is preferable for deriving the appropriate ECM algorithm.

Multiple proportionality constraints Certain linear constraints, such as those of equation (6), lead to intractable EM algorithms, depending on the parameters to estimate. This is also the case for linear constraints on the variance parameters, as we discuss in Section 3.1. For simpler situations we can define a less general setup, which we call "multiple proportionality" constraints, that result in straightforward closed-form M-steps within true EM algorithms (or ECM algorithms, in some cases). Basically, we shall designate, among m scalar component parameters (e.g., means), subsets of parameters that are related by known multiplicative constants (as in Quandt and Ramsey (1978), see Section 2). To set multiple proportionality constraints among, say, the mean parameters $\boldsymbol{\mu}$, we define a subset $J^\mu \subseteq \{1, \ldots, m\}$, known constants $\mathbf{a}^\mu = (a_j^\mu, j \in J^\mu)$, and one particular element j_0 of J^μ, such that $a_{j_0}^\mu = 1$, and $\mu_j = a_j^\mu \mu_{j_0}$ for $j \in J^\mu$. Hence, j_0 plays the arbitrary role of labelling which one of the means in $\{\mu_j, j \in J^\mu\}$ will formally be in the model parameter that is then restricted to $\boldsymbol{\theta} = \big(\boldsymbol{\lambda}, \mu_{j_0}, (\mu_j, j \notin J^\mu), \boldsymbol{\sigma}^2\big)$. For instance, a 3-component normal mixture with simple mean constraints $\mu_3 = -\mu_2$ on the second and third components is defined by $J^\mu = \{2, 3\}$ and $\mathbf{a}^\mu = (1, -1)$. Several disjoint sets J may be similarly defined if there exist distinct sets of proportionally constrained parameters.

3 Gaussian generalized EM algorithms for constrained parameters

We begin by recalling the well-known M-step for the mean and variance parameters in the normal case (see, e.g., McLachlan and Krishnan, 2008):

For simplicity of notation, we will also denote the variances v_j instead of σ_j^2, so that the jth component parameter is $\xi_j = (\mu_j, v_j)$.

M-step for $\boldsymbol{\xi} = (\boldsymbol{\mu}, \mathbf{v})$ in the normal case

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} x_i}{\sum_{i=1}^n p_{ij}^{(t)}}, \quad \text{for } j = 1, \ldots, m, \tag{7}$$

$$v_j{}^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n p_{ij}^{(t)}}, \quad \text{for } j = 1, \ldots, m. \tag{8}$$

Unfortunately, in the presence of constraints such as those in Section 2, the M-step can be rather complicated and typically it has no closed form. However, it is often easier to compute the M-step conditionally on some of the parameters. We thus propose to use an extension of the EM algorithm, known as the ECM, or Expectation-Conditional Maximization, algorithm, a class of algorithms introduced by Meng and Rubin (1993) (see also McLachlan and Krishnan, 2008, Chapter 5). An ECM algorithm replaces a complicated M-step with several computationally simpler CM-steps. Sometimes, even an ECM algorithm does not lead to tractability. We show in Section 3.3 how this problem may be overcome using yet another generalization of EM algorithms, the class of so-called Minorization-Maximization, or MM algorithms (Hunter and Lange, 2004).

The standard normal EM as well as the constrained EM, ECM, and MM algorithms that are defined in this paper are implemented in the mixtools package (Young et al. (2011) and Benaglia et al. (2009)) for the R statistical software (R Core Team, 2012). We choose here to handle only the normal case, but extensions to any parametric family with mean and/or variance parameters are often straightforward.

3.1 Multiple proportional constraints

The estimates of the parameters are straightforward in this case, for both means and variances. For constraints on the means, set as in section 2.1 a subset $J^\mu \subseteq \{1, \ldots, m\}$, constants $\mathbf{a}^\mu = (a_j^\mu, j \in J^\mu)$, and a fixed $j_0 \in J^\mu$ such that $a_{j_0}^\mu = 1$ and $\mu_j = a_j^\mu \mu_{j_0}$ for $j \in J^\mu$. The model parameter is restricted to $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \mu_{j_0}, (\mu_j, j \notin J^\mu), \mathbf{v})$. Maximization with respect to all the parameters except μ_{j_0} is unchanged, and maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with

respect to μ_{j_0} gives

$$\mu_{j_0}^{(t+1)} = \left(\sum_{i=1}^{n} \sum_{j \in J^\mu} \frac{p_{ij}^{(t)}}{v_j^{(t)}} a_j^\mu x_i \right) \left(\sum_{i=1}^{n} \sum_{j \in J^\mu} \frac{p_{ij}^{(t)}}{v_j^{(t)}} (a_j^\mu)^2 \right)^{-1}. \tag{9}$$

This update is conditional on $\mathbf{v} = \mathbf{v}^{(t)}$, so an ECM algorithm is required here as well.

Note that in the particular case of the scale model, i.e. when all μ_j's are equal, all $a_j^\mu = 1$ and the parameter is restricted to $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \mu, \mathbf{v})$. Then equation (9) reduces to

$$\mu^{(t+1)} = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_{ij}^{(t)}}{v_j^{(t)}} x_i \right) \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_{ij}^{(t)}}{v_j^{(t)}} \right)^{-1}, \tag{10}$$

which is different than the *a priori* intuitive sample mean, since it takes into account the variances of the observations coming from each component.

Constraints on the variances Define similarly a subset $J^v \subseteq \{1, \ldots, m\}$, known constants $\mathbf{a}^v = (a_j^v, j \in J^v)$, and one fixed $j_0 \in J^v$ such that $a_{j_0}^v = 1$ and $v_j = a_j^v v_{j_0}$, $j \in J^v$. The parameter is now restricted to $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\mu}, v_{j_0}, (v_j, j \notin J^v))$. As before, maximization with respect to all the parameters except v_{j_0} is unchanged, and maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ with respect to v_{j_0} gives

$$v_{j_0}^{(t+1)} = \left(\sum_{i=1}^{n} \sum_{j \in J^v} p_{ij}^{(t+1/2)} \frac{\left(x_i - \mu_j^{(t+1)} \right)^2}{a_j^v} \right) \left(\sum_{i=1}^{n} \sum_{j \in J^v} p_{ij}^{(t+1/2)} \right)^{-1}. \tag{11}$$

If we do not also have constraints on the $\boldsymbol{\mu}$ parameters, then there is no need for an ECM algorithm, since the maximization steps (5) for $\boldsymbol{\lambda}$ and (7) for $\boldsymbol{\mu}$ do not depend on $\mathbf{v}$. In this case, the intermediate E-step (15) is unnecessary and so $p_{ij}^{(t+1/2)}$ should be replaced by $p_{ij}^{(t)}$ in equation (11).

3.2 Linear constraints on the means and ECM algorithms

The expected complete-data loglikelihood is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \sum_{j=1}^{m} p_{ij}^{(t)} \left[\log \lambda_j - \frac{1}{2} \log v_j - \frac{1}{2} \frac{(x_i - \mu_j)}{v_j} \right]. \tag{12}$$

We assume that the vector of component means satisfies $\boldsymbol{\mu} = M\boldsymbol{\beta} + \mathbf{C}$ as in (6). In the ECM framework, maximization is done conditionally on the $\mathbf{v}$ parameters; i.e., we take $v_j = v_j^{(t)}$ in (12). We focus only on the portion of the CM-step in which we fix the $\mathbf{v}$ parameters and update the $\boldsymbol{\beta}$ parameters. Differentiating (12) with respect to β_ℓ for $1 \leq \ell \leq p$, then setting these derivatives equal to zero, gives, in matrix form,

$$M^\top \mathbf{d}^{(t)} = M^\top B^{(t)} \boldsymbol{\mu},$$

where $\mathbf{d}^{(t)}$ is an m-vector defined by

$$d_j^{(t)} = \frac{1}{v_j^{(t)}} \sum_{i=1}^{n} p_{ij}^{(t)} x_i$$

and $B^{(t)}$ is an $m \times m$ diagonal matrix with jth diagonal term

$$B_{jj}^{(t)} = \frac{1}{v_j^{(t)}} \sum_{i=1}^{n} p_{ij}^{(t)}.$$

Thus,

$$M^\top \mathbf{d}^{(t)} = M^\top B^{(t)} (M\boldsymbol{\beta} + \mathbf{C}), \tag{13}$$

and the update for $\boldsymbol{\beta}$ is given in closed form by

$$\boldsymbol{\beta}^{(t+1)} = \left(M^\top B^{(t)} M \right)^{-1} M^\top \left(\mathbf{d}^{(t)} - B^{(t)} \mathbf{C} \right)$$

The update for $\boldsymbol{\mu}$ at iteration $t+1$ is thus

$$\boldsymbol{\mu}^{(t+1)} = M\boldsymbol{\beta}^{(t+1)} + \mathbf{C}, \tag{14}$$

which is used in the CM-step (8) for updating $\mathbf{v}$.

A typical iteration of an ECM algorithm consists of multiple sub-iterations: The CM-steps maximize first with respect to one subset of the parameters, then another, and so on until the full parameter vector has been updated. In between each pair of CM-steps is another E-step. Equation (14), for example, only describes a single CM-step, which updates the $\boldsymbol{\mu}$ parameters. We may also use equation 5 to update the $\boldsymbol{\lambda}$ parameters in the same CM-step, since the $\boldsymbol{\lambda}$ update does not affect the $\boldsymbol{\mu}$ parameters. However, after updating $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ using equations (5) and (14), it is necessary to interject

a second E-step before updating $\boldsymbol{\sigma}$. This extra E-step consists of defining $Q(\boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{v} \mid \boldsymbol{\lambda}^{(t+1)}, \boldsymbol{\mu}^{(t+1)}, \mathbf{v}^{(t)})$, as in equation (3), and

$$p_{ij}^{(t+1/2)} = \frac{\lambda_j^{(t+1)} f\left(x_i; \mu_j^{(t+1)}, v_j^{(t)}\right)}{\sum_{r=1}^{m} \lambda_r^{(t+1)} f\left(x_i; \mu_r^{(t+1)}, v_r^{(t)}\right)} \tag{15}$$

before updating the $\mathbf{v}$ parameters in the next CM-step. If there are no constraints on the $\mathbf{v}$ parameters, one may simply use equation (8) with $p_{ij}^{(t+1/2)}$ in place of $p_{ij}^{(t)}$. We discuss the case where constraints are placed on $\mathbf{v}$ in Sections 3.1 and 3.3.

3.3 Linear constraints on the variances and MM algorithms

If linear constraints such as those in Section 3.2 are desired for the variance parameters, the approach used in that section will not work for constructing a CM-step because there is no closed-form maximizer of the Q function with respect to the constrained $\mathbf{v}$ parameters. Of course, one could use a numerical maximization technique in this case, but here we demonstrate an alternative that admits a closed form.

We first reparameterize by letting $\pi_j = 1/v_j$ for $1 \leq j \leq m$. Then, assume that the constraints are given by $\boldsymbol{\pi} = A\boldsymbol{\gamma}$, where A is a known $m \times q$ matrix with nonnegative entries and $\boldsymbol{\gamma}$ is the q-vector of the (unknown) parameters of interest, with $q \leq m$. The parameter $\boldsymbol{\gamma}$ will be guaranteed to have positive coordinates by the algorithm we derive as equation (17).

Assuming that $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ have already been updated in the first half of an ECM iteration, we wish to calculate the expected loglikelihood for the complete data, $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1/2)})$. The part of this function that depends on the $\boldsymbol{\pi}$ parameters is

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t+1/2)} \log \pi_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t+1/2)} \pi_j \left(x_i - \mu_j^{(t+1)}\right)^2, \tag{16}$$

which does not admit a closed-form maximizer. Therefore, we shall rely instead on a so-called MM, or minorization-maximization, algorithm. These algorithms have a long history in the statistical literature, a history that far predates the use of the initials MM; details may be found in Hunter and Lange (2004).

Since $\boldsymbol{\pi} = A\boldsymbol{\gamma}$, we may express the jth component of $\boldsymbol{\pi}$ as

$$\pi_j = A_j\boldsymbol{\gamma} = \sum_{k=1}^{q} A_{jk}\gamma_k,$$

where A_j is the jth row of A. The essential MM idea is this: Since the logarithm function is a concave function, we may use inequality (10) in Hunter and Lange (2004) to prove that

$$\log \pi_j \geq \sum_{k=1}^{q} \frac{A_{jk}\gamma_k^{(t)}}{\pi_j^{(t)}} \log\left(\frac{\pi_j^{(t)}\gamma_k}{\gamma_k^{(t)}}\right),$$

with equality when $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t)}$. We conclude that the function defined by

$$\begin{aligned}\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} p_{ij}^{(t+1/2)} \frac{A_{jk}\gamma_k^{(t)}}{\pi_j^{(t)}} \log\left(\frac{\pi_j^{(t)}\gamma_k}{\gamma_k^{(t)}}\right) \\ -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{k=1}^{q} p_{ij}^{(t+1/2)} A_{jk}\gamma_k (x_i - \mu_j^{(t+1)})^2\end{aligned}$$

has the property that it is bounded above by (16), with equality when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t+1/2)}$. Thus, maximizing this function will result in an increase in the value of $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t+1/2)})$, which in turn means that this MM algorithm is an example of GEM in the sense of Dempster et al. (1977) and it increases the observed data likelihood.

Setting the partial derivatives with respect to γ_ℓ equal to zero for $1 \leq \ell \leq q$, we obtain

$$\gamma_\ell^{(t+1)} = \gamma_\ell^{(t)} \left[\frac{\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{p_{ij}^{(t+1/2)} A_{j\ell}}{\pi_j^{(t)}}\right)}{\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t+1/2)} A_{j\ell}\left(x_i - \mu_j^{(t+1)}\right)^2} \right]. \tag{17}$$

Since A has nonnegative entries, the requirement that $\boldsymbol{\gamma}^{(t+1)} \geq 0$ is automatically enforced by this algorithm.

4 Hypothesis testing for constraints

Testing and model selection methods generally require estimation algorithms for their implementation, and the algorithms we present here can serve this purpose. The general model parameter space in the present case of univariate Gaussian mixture is

$$\Theta = \{(\boldsymbol{\lambda}, \boldsymbol{\mu}, \mathbf{v}) \in [0,1]^{m-1} \times \mathbb{R}^m \times (\mathbb{R}_*^+)^m\},$$

with dimension $d = \dim(\Theta) = 3m - 1$, whereas the parameter space under, say, linear constraints for both means and variances is

$$\Theta_0 = \{(\boldsymbol{\lambda}, M\boldsymbol{\beta} + C, A\boldsymbol{\gamma}), \boldsymbol{\lambda} \in [0,1]^{m-1}, \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\gamma} \in (\mathbb{R}_*^+)^q\}$$

with dimension $d_0 = \dim(\Theta_0) = m + p + q - 1$, $p < m$, $q < m$ (since the cases $p = q = m$ are merely reparametrization).

Since the constrained model is a special case of (is nested in) the unconstrained one, a null hypothesis such as "the constraints hold", that is H_0 : "$\boldsymbol{\theta} \in \Theta_0$" can be tested by a standard Likelihood Ratio Test (LRT), with statistic

$$\Lambda_n = 2 \log \frac{\sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}; \mathbf{x})}{\sup_{\boldsymbol{\theta}_0 \in \Theta_0} L(\boldsymbol{\theta}_0; \mathbf{x})}, \tag{18}$$

where $L(\boldsymbol{\theta}; \mathbf{x})$ is the observed likelihood of the model, Θ_0 is the parameter space of the null model satisfying the constraint, and Θ is the full model parameter space. Denote $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_0$ the unconstrained and constrained (under H_0) maximum likelihood estimators of $\boldsymbol{\theta}$, respectively. The plain EM and the ECM or EC-MM algorithms presented in this paper can be used at this point to estimate $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_0$, respectively. Then $\Lambda_n = 2(\ell(\hat{\boldsymbol{\theta}}; \mathbf{x}) - \ell(\hat{\boldsymbol{\theta}}_0; \mathbf{x}))$ is asymptotically $\chi^2(d - d_0)$-distributed, where $\ell()$ is the log-likelihood of the model (see, e.g., van der Vaart, 1998).

5 Examples and simulation studies

Here we compare plain (i.e., without constraints) EM algorithms against our constrained ECM and "EC-MM" algorithms. All examples are run using version 1.0 of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Core Team, 2012), which is publicly available on the Comprehensive R Archive Network (CRAN) at `cran.r-project.org`.

We have ignored the "label-switching" issue up to now. This issue arises because the particular ordering of the subscripts $j = 1, \ldots, m$ in equation (1) is arbitrary: A rearrangement of these subscripts gives exactly the same density function even though technically the elements of the parameter vector $\boldsymbol{\theta}$ have been permuted. Thus, this "label-switching" possibility destroys the usual parameter identifiability assumption that underlies statistical inference, whereby the distribution of the data uniquely determines the parameter values. In practice, since we are only concerned with finding a single maximum likelihood estimator of $\boldsymbol{\theta}$ here, this lack of identifiability causes no problems as long as we keep in mind that the best we can do is to estimate the parameters up to a permutation of the labels. However, in a

simulation study based on Monte-Carlo replications, this issue can lead to flawed average estimates because there is no guarantee that only estimates from the "same" component are averaged together. For a fuller account of the label-switching issue, see McLachlan and Peel (2000).

As is typically the case for EM-related algorithms, the choice of initial starting parameter values is important. In these tests, we started the algorithms from the true parameter values to prevent label-switching as much as possible. Of course, the true values are not known in practice; however, one would typically use multiple different random starting values, then take as a final estimator the solution yielding the highest likelihood. This result typically is the same as what one observes in a simulation study starting at the true parameter values.

We declare convergence using a simplistic lack-of-progress criterion that considers only the change in log-likelihood value from one iteration to the next, stopping the algorithm when the difference is smaller than 10^{-8}. As McNicholas et al. (2010) point out, a criterion based on Aitken acceleration may also be used here; indeed, these authors argue that the more sophisticated Aitken criterion is superior. However, we do not consider this question in the current article.

5.1 Examples of EM with constrained parameters

We consider as an example requiring some parameter constraints, a model from Thomas et al. (2011) used for assessing reliability of repeated measurements in psychometrics. Briefly, these authors end up with two possible $m = 3$ components Gaussian mixtures with constained parameters.

Parallel test models This first model is expressed by

$$\lambda_1 \mathcal{N}(0, v_1) + \lambda_2 \mathcal{N}(\mu_2, v_1 + \sigma^2) + \lambda_3 \mathcal{N}(-\mu_2, v_1 + \sigma^2), \tag{19}$$

i.e. a 3-component normal mixture with constraints

$$\mu_1 = 0, \quad \mu_3 = -\mu_2, \quad v_3 = v_2 > v_1. \tag{20}$$

The rather simple constraint for the means can be handled by our multiple proportionality scheme, so that we will use the corresponding algorithm and code in this case. With the reparametrization $\pi_j = 1/v_j$ from Section 3.3, the constraints on the variances are expressed by

$$\boldsymbol{\pi} = A\boldsymbol{\gamma}, \quad \text{with } A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} \tag{21}$$

and the inequality constraint in (20) is satisfied as well since the γ_j's are positive, $\pi_2 = \pi_3 = \gamma_1 < \pi_1 = \gamma_1 + \gamma_2$.

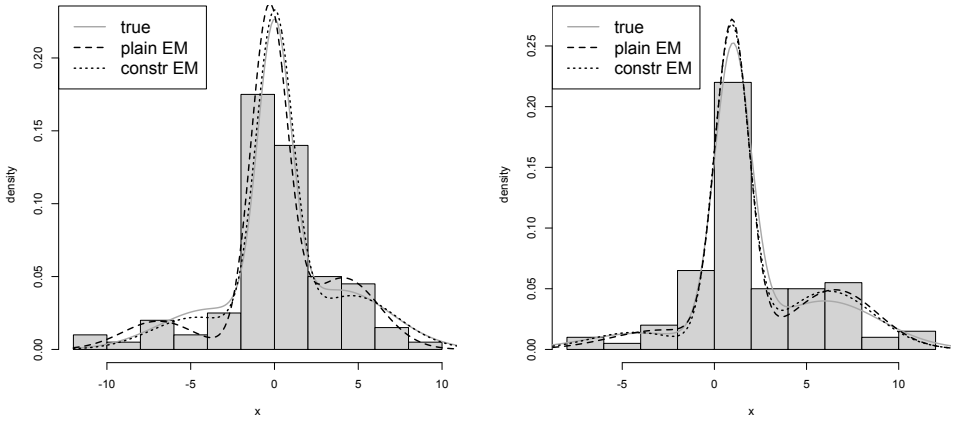

Figure 1: *True and estimated mixture densities for plain Gaussian EM and constrained EC-MM algorithms on a sample of size* $n = 100$ *for the parallel test synthetic model (left) and the tau equivalent model (right).*

We define such a synthetic parallel test model with: $\boldsymbol{\lambda} = (0.5, 0.3, 0.2)^\top$, $\boldsymbol{\mu} = (0, 4, -4)^\top$, and $\boldsymbol{\sigma} = (1, 3, 3)^\top$. The true parameters have been chosen to assure a unimodal, overlapping mixture density, as depicted in Fig. 1 (left). In this situation, the constraints on the parameter space play a more prominent role in the EM estimation than if the components were well-separated. Fig. 1 (left) also shows some EM and constrained EC-MM estimates on a small sample ($n = 100$), illustrating the good behavior of the constrained version. Mean squared errors are provided in Table 1 to compare the behavior of the plain (unconstrained) EM and the constrained version of the Gaussian EC-MM algorithm on the basis of 300 replications. It is clear for both small ($n = 100$) and large ($n = 1000$) sample sizes that constraining the parameter space helps both in terms of bias and MSE.

We also investigate the frequency of label-switching in our algorithm. For $n = 100$ and the unconstrained EM version, we actually observed 15% of the 300 replications for which $\hat{\lambda}_1 < \hat{\lambda}_2$, which suggests that label-switching has occurred despite our choice of starting values. However, further investigation reveals that the mean and variance parameter estimates did *not* appear to be switched in these cases; thus, they evidently merely represent poor estimates of λ_1 and λ_2. For the 18% of replications of the constrained versions for

which we got $\hat{\lambda}_1 < \hat{\lambda}_2$, this was even more obvious due to the constraints on the means and variances that give less flexibility in the parameters. For $n = 1000$ we observed similar behavior among the 4% of the plain EM replications for which $\hat{\lambda}_1 < \hat{\lambda}_2$, and no such inversion for the constrained version.

component		$j=1$		$j=2$		$j=3$	
algorithm		U	C	U	C	U	C
$n = 100$	$\boldsymbol{\lambda}$	0.0196	0.0182	0.0266	0.0142	0.0244	0.0069
	$\boldsymbol{\mu}$	0.0499		3.9226	1.89	6.2823	1.89
	$\boldsymbol{\sigma}$	0.0828	0.068	1.1505	0.641	1.6178	0.641
$n = 1000$	$\boldsymbol{\lambda}$	0.0022	0.0020	0.0052	0.00095	0.0045	0.0005
	$\boldsymbol{\mu}$	0.0048		0.7055	0.180	1.2222	0.180
	$\boldsymbol{\sigma}$	0.00521	0.0051	0.16407	0.06179	0.27833	0.0618

Table 1: Estimated Mean Squared Errors from 300 replications of Unconstrained Gaussian EM (U) and Constrained EC-MM (C) algorithms started from the true parameters, for the parallel test synthetic model, and two sample sizes n.

Tau equivalent model Thomas et al. (2011) also consider a more general model, that is a 3-component normal mixture satisfying the same constraints on the variances as before and, for the mean, a linear constraint that can be expressed as in (6) with $\mathbf{C} = \mathbf{0}$, $\mu_1 = \beta_1$, $\mu_2 = \beta_1 + \beta_2$, and $\mu_3 = \beta_1 - \beta_2$. We simulate a synthetic model with true parameters $\boldsymbol{\lambda} = (0.6, 0.3, 0.1)$ and $\boldsymbol{\beta} = (1, 5)$ so that $\boldsymbol{\mu} = (1, 6, -4)$, and $\boldsymbol{\sigma} = (1, 3, 3)$. The three components of this example are more separated than those of the previous model, but the weight of the "unstable negative individuals" is assumed to be smaller (10% of the population), and hence this component is more difficult to estimate precisely from a small sample.

The true density, together with sample estimators, is depicted in the right panel of Fig. 1. Results in terms of the MSE's are in Table 2. As for the parallel test model, we did observe some inversions of the λ estimates (i.e., where $\hat{\lambda}_1 < \hat{\lambda}_2$) though there was no label switching here. These inversions only occurred for the $n = 100$ case.

5.2 Examples of hypothesis testing

We applied the Likelihood Ratio test detailed in Section 4 for H_0 :"the constraints hold" to each of the above examples, for which the full model

component		$j=1$		$j=2$		$j=3$	
algorithm		U	C	U	C	U	C
	$\boldsymbol{\lambda}$	0.0125	0.00941	0.0127	0.0093	0.0116	0.002
$n=100$	$\boldsymbol{\mu}$	0.0702	0.0308	2.5847	1.6136	7.0203	1.5063
	$\boldsymbol{\sigma}$	0.0959	0.0411	0.8583	0.5297	2.2135	0.5297
	$\boldsymbol{\lambda}$	0.0009	0.0008	0.00129	0.0006	0.0007	0.0001
$n=1000$	$\boldsymbol{\mu}$	0.0025	0.0022	0.2311	0.1216	1.0290	0.1100
	$\boldsymbol{\sigma}$	0.0026	0.0024	0.0858	0.0442	0.2957	0.0442

Table 2: Estimated Mean Squared Errors from 300 replications of Unconstrained Gaussian EM (U) and Constrained EC-MM (C) algorithms started from the true parameters, for the tau equivalent synthetic model, and two sample sizes n.

has $3m-1=8$ parameters. For the Parallel test model, Θ_0 is given by (19) and (20), and the LRT statistics is theoretically asymptotically $\chi^2(3)$-distributed, since the constrained model reduces the parameters to 2 proportions, $p=1$ mean (μ_2) and $q=2$ (inverse) variances. For the tau equivalent model, the LRT statistics is asymptotically $\chi^2(2)$-distributed since $p=q=2$.

The parameters of the simulated models under H_0 has been set as in the preceding section, and for each of the two models, two alternative hypothesis H_1^j, $j=1,2$ have been simulated, with same $\boldsymbol{\lambda}$ and means and variances given in Table 5.2 below for all the situations. Empirical levels and powers for a nominal 5% level test, both based on 500 replications, asymptotic distributions and increasing sample sizes are displayed in Fig. 2.

		$\boldsymbol{\mu}$			$\mathbf{v}$		
	H_0	0	4	-4	1	9	9
parallel	H_1^1	0.5	5	-4	1	8	9
	H_1^2	0.2	5	-4	1	8	9
	H_0	1	6	-4	1	9	9
tau eq.	H_1^1	3	7	-4	2	8	9
	H_1^2	0	6	-5	1	8	9

Table 3: Parameters for the simulated models under H_0 and two choices for the alternative hypothesis, for each of the two models.

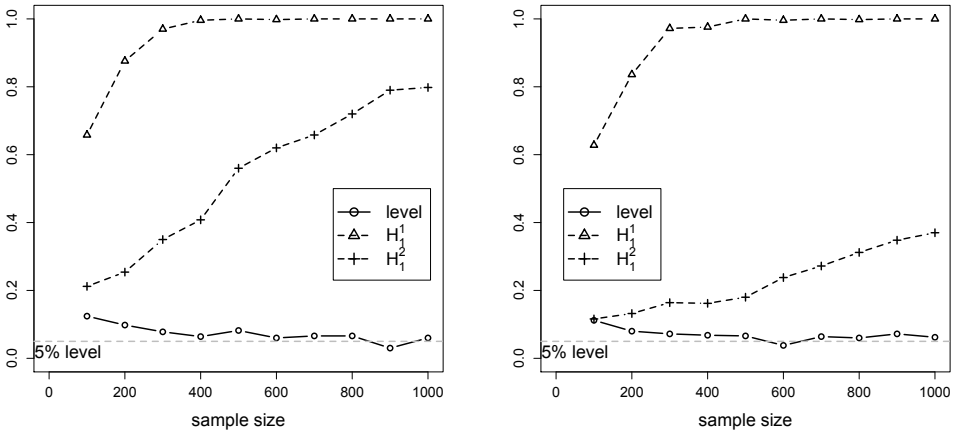

Figure 2: *LRT empirical levels (solid) and powers (dashed) for a nominal 5% level for the parallel test model (left) and tau equivalent (right). Alternative hypothesis* H_1^j, $j = 1, 2$ *are as given by Table 5.2. Results are based on 500 replications and asymptotic distribution* $\chi^2(3)$ *(left) and* $\chi^2(2)$ *(right).*

6 Discussion

The algorithms we propose in this article extend the well-known EM algorithm for finite mixture models to the case with various linear constraints on the space of parameters. We show that in the presence of such constraints, the M-step typically has no closed form, but ECM and sometimes "EC-MM" (i.e., conditional MM-steps) extensions of the EM algorithm with closed-form implementations can be defined. Note that all the extensions we develop here share with genuine EM algorithms the same essential ascent property of the observed likelihood function.

Our simulations show that constraints on the parameter space can improve the estimation in terms of mean squared error relative to estimates calculated without assuming constraints. Hypothesing testing for null hypothesis such as "constraints hold" can also been tested by a standard Likelihood Ratio Test (LRT) and we have shown the behavior of such test in some examples.

We also choose in this article to handle only the mixture-of-Gaussian-components case, but extensions to any parametric family with mean and/or variance parameters, or more generally to components with likelihood leading to closed-form maximization, should allow for similar ideas.

Finally, we reiterate that the algorithms presented in this article are implemented in the R package called mixtools (Benaglia et al., 2009) for the R statistical software (R Core Team, 2012), which is publicly available on the Comprehensive R Archive Network (CRAN) at `cran.r-project.org`.

References

Banfield, J. D. and Raftery, A. E. (1993). Model-based gaussian and non-gaussian clustering. *Biometrics*, 49(3):803–821.

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

Bensmail, H. and Celeux, G. (1996). Regularized gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91(436):1743–1748.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Hathaway, R. J. (1985). A constrained formulation of maximum-likelihood estimation for normal mixture distribution. *Annals of Statistics*, 13:795–800.

Hathaway, R. J. (1986). A constrained EM algorithm for univariate normal mixtures. *J. Statist. Comput. Simul.*, 23:211–230.

Hunter, D. R. and Lange, K. (2004). A tutorial on MM algorithms. *The American Statistician*, 58:30–37.

Kim, D. K. and Taylor, J. M. G. (1995). The restricted EM algorithm for maximum likelihood estimation under linear restrictions on the parameters. *J. Amer. Statist. Assoc.*, 90:708–716.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

McNicholas, P. D., Murphy, T. B., McDaid, A. F., and Frost, D. (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis*, 54(3):711–723.

Meng, X.-L. and Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, 80:267–278.

Mooijaart, A. and van der Heijden, P. G. M. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika*, 2:261–269.

Nettleton, D. (1999). Convergence properties of the EM algorithm in constrained parameter spaces. *Canadian Jour. Statist.*, 27:639–648.

Quandt, R. E. and Ramsey, J. B. (1978). Estimating mixtures of normal distributions and switching regressions. *J. Amer. Statist. Assoc.*, 73:730–738.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Shi, N. Z., Zheng, S. R., and Guo, J. (2005). The restricted EM algorithm under inequality restrictions on the parameters. *J. Mult. Analysis*, 92:53–76.

Thomas, H., Lohaus, A., and Domsch, H. (2011). Extensions of reliability theory. In Hunter, D. R., Richards, D. St. P., and Rosenberger, J. L., editors, *Nonparametric Statistics and Mixture Models: A Festschrift in Honor of Thomas P. Hettmansperger*, pages 309–316, Singapore. World Scientific.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Young, D. S., Benaglia, T., Chauveau, D., Elmore, R. T., Hettmansperger, T. P., Hunter, D. R., Thomas, H., and Xuan, F. (2011). mixtools: Tools for mixture models. R package version 1.0.