ECM and MM algorithms for mixtures with constrained parameters

Didier Chauveau, David R. Hunter

ECM and MM algorithms for mixtures with constrained parameters

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

[1]Université d'Orléans and CNRS UMR 6628, France
[2]Pennsylvania State University, PA, USA

September 20, 2011

Abstract: EM algorithms for obtaining maximum likelihood estimates of parameters in finite mixture models are well-known, and normal mixtures are the most commonly used in this category. In fact, certain types of constraints on the parameter space, such as the equality of variance assumption, lead to well-known EM algorithms. After briefly summarizing these well-known results, we then consider the problem of more general constraints on the parameter space for finite mixtures of normal components. Surprisingly, this simple extension has not been explored in the literature. Here, we show how the MLE problem succumbs to an EM generalization known as an ECM algorithm. With 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 the problem of parameter estimation in finite mixtures of normal components in the presence of equality or linear constraints on the parameters. We provide software that implements these algorithms.

Keywords: generalized EM algorithms, ECM algorithms, MM algorithms, finite mixture.

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 and since this is the case for the motivating example, 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); here we only describe the finite-mixture case.

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 $\boldsymbol{\theta}^{(t+1)} = \operatorname{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$.

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 constraints on the space of parameters in the finite mixture of normal distributions. In so doing, this section describes a very commonly used constraint and a more unusual constraint that is motivated by a question from psychometrics. Section 3 shows how an EM algorithm may be modified to handle a general class of constraints. Section 4 revisits the motivating example of Section 2 by presenting a simulation study in the spirit of the psychology dataset, and Section 5 summarizes the article.

2 Constraints on the parameter space

In the case of normal component densities, which will be the sole topic of the remainder of this article, 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; we leave this exercise to the interested reader.

We now consider an example that leads to different parameter constraints. Thomas et al. (2011) describe an extension to the theory of reliability of repeated measurements in psychometrics. Here, "reliability" has nothing to do with statistical survival data analysis; it refers to the study of aggreement between similar psychometric tests given on two or more occasions to the same sample of individuals. In a typical application, two score results are recorded on each of a sample of n individuals from a population of study, where the observed value $Y_{k,i}$ for individual i on test k is assumed to be the sum of an unobservable true score (T) and some independent measurement error (E):

$$Y_{k,i} = T_{k,i} + E_{k,i}, \quad k = 1, 2, \quad i = 1, \ldots, n.$$

A typical assumption is that the true scores for an individual are perfectly correlated. More precisely, the *tau equivalent* model stipulates that there exists a constant α such that $T_{2i} = T_{1i} + \alpha$ for all i, whereas the *parallel test* model is the special case of the tau equivalent model in which $\alpha = 0$. The *reliability coefficient* is simply the correlation $\rho(Y_1, Y_2) = \sigma_T^2/(\sigma_T^2 + \sigma_E^2)$, where σ_T^2 and σ_E^2 are the variances of the true scores and errors.

Thomas et al. (2011) propose a 3-component mixture model in which one component consists of *stable* individuals who follow the traditional parallel test model (or the tau equivalent model, if $\alpha \neq 0$ is desired). By contrast, the other two components consist of *unstable* individuals. In terms of the observed differences

$$D_i = Y_{1,i} - Y_{2,i} = (T_{1,i} - T_{2,i}) + (E_{1,i} - E_{2,i}), \tag{6}$$

we postulate the mixture model

$$D_i \sim \lambda_1 \mathcal{N}(0, 2\sigma_E^2) + \lambda_2 \mathcal{N}(\mu_\delta, \sigma^2) + \lambda_3 \mathcal{N}(-\mu_\delta, \sigma^2). \tag{7}$$

Here, we define $\Delta_i = T_{1,i} - T_{2,i}$ so that the individual distributions of the T_k are not of interest; only the differences Δ and errors E are considered. It is always assumed there that $E \sim \mathcal{N}(0, \sigma_E^2)$ and that Δ is normal, with nonzero mean, for the unstable individuals. Because instability is assumed to result in both positive and negative effects, at least $m = 3$ components are needed to describe the population. The authors also assume that the two subgroups of the unstable population have means of equal magnitude but opposite signs and that their normal distributions have the same variance,

given by $\sigma^2 = 2\sigma_E^2 + \sigma_\Delta^2$. All of these facts except the necessary inequality $\sigma^2 \geq \sigma_E^2$ are summarized by equation (7), and the parallel test model from Thomas et al. (2011) amounts to a 3-component normal mixture with the following constraints:

$$\mu_1 = 0, \quad \mu_3 = -\mu_2, \quad \sigma_3 = \sigma_2 > \sigma_1. \tag{8}$$

Thomas et al. (2011) also handle the *tau equivalent model*, where $T_2 - T_1 = \alpha \neq 0$. The mixture model then satisfies the following constraints on the mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$:

$$\mu_j = \begin{cases} \alpha & \text{if } j = 1 \\ \alpha + \delta & \text{if } j = 2 \\ \alpha - \delta & \text{if } j = 3. \end{cases} \tag{9}$$

In what follows, we will deal with both sets of constraints. Each of these two models includes an inequality constraint on the variances, namely $\sigma_2 = \sigma_3 > \sigma_1$. We demonstrate in Section 3.4 that this constraint may be reformulated using equality constraints so that it may be handled using the techniques of this article.

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. It seems that 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 Taylord (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.

2.1 Two definitions of constraints

Let us consider the mean constraints in equations (8) and (9). One way of denoting these constraints generally is to write

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_m \end{pmatrix} = M\boldsymbol{\beta} + \mathbf{C} = M \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} C_1 \\ \vdots \\ C_m \end{pmatrix} \tag{10}$$

for some known $m \times p$ matrix M, some unknown p-vector $\boldsymbol{\beta}$, with $p \leq m$, and a known fixed m-vector of constants $\mathbf{C}$. In the parallel test model (8), and with the notation of the previous section, we have $p = 1$ and

$$\beta = \mu_2, \quad M = \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

For the tau equivalent model (9), we have

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \delta \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

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 3.2 why this is incorrect.

Multiple proportionality constraints Certain linear constraints, such as those of equation (10), 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 3.3. For such situations we can define a less general setup, which we call "multiple proportionality" constraints, that result in closed-form M-steps within true EM algorithms (or ECM algorithms, in some cases). Basically, we shall designate, among the m scalar component parameters (e.g., means), subsets of parameters that are related by known multiplicative constants.

Let us consider a multiple proportionality constraint among the mean parameters $\boldsymbol{\mu}$. We define a subset $J^\mu \subseteq \{1, \ldots, m\}$ and known constants $\mathbf{a}^\mu = (a_j^\mu, j \in J^\mu)$, such that

$$\text{for } j_0 \in J^\mu, a_{j_0}^\mu = 1, \quad \text{and} \quad \mu_j = a_j^\mu \mu_{j_0}, \quad j \in J^\mu.$$

Here, j_0 plays the arbitrary role of labelling which one of the mean parameters will be estimated in the EM algorithm. For instance, the constraint on the means in the parallel test model (8) is $J^\mu = \{2, 3\}$ 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

$$\mu_j^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} x_i}{\sum_{i=1}^n p_{ij}^{(t)}}, \quad \text{for } j = 1, \ldots, m, \tag{11}$$

$$v_j{}^{(t+1)} = \frac{\sum_{i=1}^n p_{ij}^{(t)} (x_i - \mu_j^{(t+1)})^2}{\sum_{i=1}^n p_{ij}^{(t)}}, \quad \text{for } j = 1, \ldots, m. \tag{12}$$

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 3.4 how this problem may be overcome using yet another generalization of EM algorithms, the class of so-called 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. (2009) and Benaglia et al. (2009)) for the R statistical software (R Development Core Team, 2010). 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 Linear constraints on the means and ECM algorithms

The expected complete-data loglikelihood is

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t)} \left[\log \lambda_j - \frac{1}{2}\log v_j - \frac{1}{2}\frac{(x_i-\mu_j)}{v_j}\right]. \tag{13}$$

We assume that the vector of component means satisfies $\boldsymbol{\mu} = M\boldsymbol{\beta} + \mathbf{C}$ as in (10). In the ECM framework, maximization is done conditionally on the $\mathbf{v}$ parameters; i.e., we take $v_j = v_j^{(t)}$ in (13). 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 (13) with respect to β_ℓ for $1 \le \ell \le p$, 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,

$$M^\top \mathbf{d}^{(t)} = M^\top B^{(t)} (M\boldsymbol{\beta} + \mathbf{C}), \tag{14}$$

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

$$\boldsymbol{\mu}^{(t+1)} = M\boldsymbol{\beta}^{(t+1)} + \mathbf{C}, \tag{15}$$

which is used in the CM-step (12) 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 (15), 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 (15), 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

$$p_{ij}^{(t+1/2)} = \frac{\lambda_j^{(t+1)} f\left(x_i; \mu_j^{(t+1)}, v_j^{(t)}\right)}{\sum_{r=1}^{m} \lambda_r^{(t+1)} f\left(x_i; \mu_r^{(t+1)}, v_r^{(t)}\right)} \tag{16}$$

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 (12) 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 3.3 and 3.4.

3.2 Multiple proportional constraints on the means

The estimates of the means are simplest in the case of multiple proportionality constraints, stipulated as in section 2.1 by a subset (or several subsets) $J^{\mu} \subseteq \{1, \ldots, m\}$, constants $\mathbf{a}^{\mu} = (a_j^{\mu}, j \in J^{\mu})$, and $j_0 \in J^{\mu}$ 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

$$\mu_{j_0}^{(t+1)} = \left(\sum_{i=1}^{n} \sum_{j \in J^{\mu}} \frac{p_{ij}^{(t)}}{v_j^{(t)}} a_j^{\mu} x_i \right) \left(\sum_{i=1}^{n} \sum_{j \in J^{\mu}} \frac{p_{ij}^{(t)}}{v_j^{(t)}} (a_j^{\mu})^2 \right)^{-1} . \tag{17}$$

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

A particular case: the scale model In this case all μ_j are equal, which means that all $a_j^\mu = 1$ and the parameter is restricted to $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \mu, \mathbf{v})$. Thus, equation (17) reduces to

$$\mu^{(t+1)} = \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_{ij}^{(t)}}{v_j^{(t)}} x_i \right) \left(\sum_{i=1}^{n} \sum_{j=1}^{m} \frac{p_{ij}^{(t)}}{v_j^{(t)}} \right)^{-1}, \tag{18}$$

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

3.3 Multiple proportional constraints on the variances

Constraints on the variances are simplest in the case of multiple proportionality, so we start with this situation. As in 3.2, we assume a set of multiple proportionality constraints defined by a subset $J^v \subseteq \{1, \ldots, m\}$, known constants $\mathbf{a}^v = (a_j^v, j \in J^v)$, and $j_0 \in J^v$ 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

$$v_{j_0}^{(t+1)} = \left(\sum_{i=1}^{n} \sum_{j \in J^v} p_{ij}^{(t+1/2)} \frac{\left(x_i - \mu_j^{(t+1)} \right)^2}{a_j^v} \right) \left(\sum_{i=1}^{n} \sum_{j \in J^v} p_{ij}^{(t+1/2)} \right)^{-1}. \tag{19}$$

If in this case 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 (11) for $\boldsymbol{\mu}$ do not depend on $\mathbf{v}$. In this case, the intermediate E-step (16) is unnecessary and so $p_{ij}^{(t+1/2)}$ should be replaced by $p_{ij}^{(t)}$ in equation (19).

3.4 Linear constraints on the variances and MM algorithms

If constraints such as those in Section 3.1 exist 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 $q \leq m$.

For instance, we can define in this setup the constraint on the variances needed for both the parallel test and tau equivalent models of Section 2, which include the inequality constraint $v_2 = v_3 > v_1$, using

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \quad \boldsymbol{\gamma} = \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

so that $\pi_2 = \pi_3 = \gamma_1 < \pi_1 = \gamma_1 + \gamma_2$, where $\boldsymbol{\gamma}$ will be guaranteed to have positive coordinates by the algorithm we derive as equation (21).

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

$$\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t+1/2)} \log \pi_j - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t+1/2)} \pi_j \left(x_i - \mu_j^{(t+1)}\right)^2, \tag{20}$$

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

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

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

has the property that it is bounded above by (20),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 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

$$\gamma_\ell^{(t+1)} = \gamma_\ell^{(t)}\left[\frac{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\frac{p_{ij}^{(t+1/2)}A_{j\ell}}{\pi_j^{(t)}}\right)}{\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}^{(t+1/2)}A_{j\ell}\left(x_i-\mu_j^{(t+1)}\right)^2}\right]. \tag{21}$$

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

4 Simulation studies

Here we compare plain (i.e., without constraints) EM algorithms against our constrained ECM and "EC-MM" versions. All examples are run using version 1.0 of the mixtools package (Benaglia et al., 2009) for the R statistical software (R Development Core Team, 2010), which is publicly available on the Comprehensive R Archive Network (CRAN) at `cran.r-project.org`. We focus here on synthetic and actual data corresponding to the motivating situation detailed in Section 2, namely, reliability of repeated mesurements in psychometrics. The specific study that motivated this algorithm involves infant habituation. Thomas et al. (2011) introduce and analyze this study using our constrained EM algorithm from the mixtools package. Though they do not enforce the inequality constraint $\sigma_3 = \sigma_2 > \sigma_1$, as we describe in Section 3.4, their solution does result in $\hat{\sigma}_3 = \hat{\sigma}_2 > \hat{\sigma}_1$. Thus, we refer the interested reader to Thomas et al. (2011) for a full account of this example, including R code to perform the analyses; this dataset can be loaded within mixtools by using the command `data(Habituationdata)`. Instead, we focus here on simulation studies.

Remark: 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).

4.1 Parallel test models

The parallel test model of Section 2 corresponds to a 3-component Gaussian mixture with simple (multiplicative and fixed) constraints on the means and linear constraints on the variances; see Section 2 and Thomas et al. (2011). In particular,

$$\mu_1 = 0, \quad \mu_3 = -\mu_2, \quad \sigma_3 = \sigma_2 > \sigma_1,$$

the constraints on $\boldsymbol{\sigma}$ being expressed in terms of the inverse variances $\boldsymbol{\pi}$ by

$$\boldsymbol{\pi} = A\boldsymbol{\gamma}, \quad A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}, \tag{22}$$

where the γ_i are positive by construction; see Equation (21). We choose first a synthetic model: $\boldsymbol{\lambda} = (0.5, 0.3, 0.2)^\top$, $\boldsymbol{\mu} = (0, 4, -4)^\top$, 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. Boxplots in Fig. 2 and mean squared errors in Table 1 are provided 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.

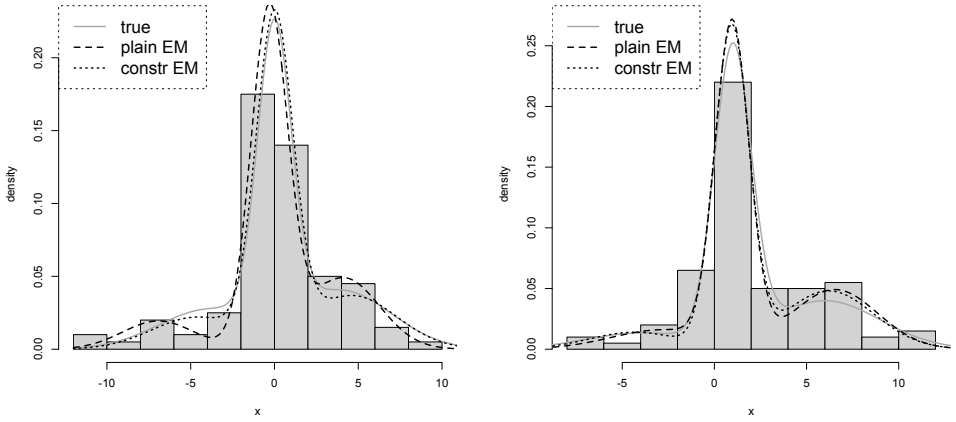

Figure 1: *True and estimated mixture densities for plain Gaussian EM and constrained EC-MM algorithms on a sample of size* $n = 100$ *for the parallel test synthetic model (left) and the tau equivalent model (right).*

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. 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.

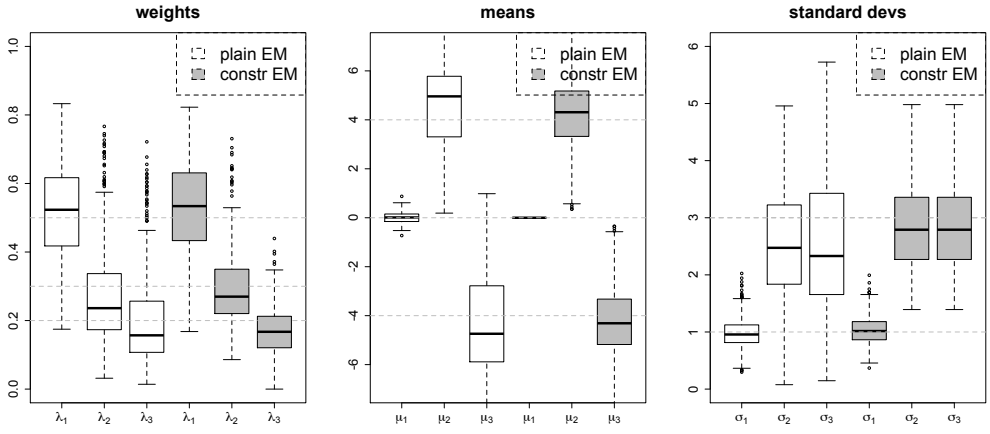

Figure 2: *Boxplots of parameter estimates from* 300 *replications of plain unconstrained EM and constrained EC-MM algorithms on samples of size* $n = 100$ *for the parallel test synthetic model; horizontal dashed lines are true values.*

4.2 Tau equivalent model

The tau equivalent model described in Section 2 corresponds to $\boldsymbol{\mu} = M\boldsymbol{\beta}$ and $\boldsymbol{\pi} = A\boldsymbol{\gamma}$, where

$$M = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & -1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad A = \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 0 \end{pmatrix}.$$

That is, the linear variance constraints coincide with those of the parallel test model (22), whereas the (linear) mean constraints are new. We simulate a 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 example in Section 4.1, 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 here are very similar to those of the parallel test model of Section 4.1, so we display only the MSEs in Table 2 for the sake of brevity. 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

		plain Gaussian EM			constrained EC-MM		
n		$j=1$	$j=2$	$j=3$	$j=1$	$j=2$	$j=3$
	$\boldsymbol{\lambda}$	0.0196	0.0266	0.0244	0.0182	0.0142	0.0069
100	$\boldsymbol{\mu}$	0.0499	3.9226	6.2823		1.89	1.89
	$\boldsymbol{\sigma}$	0.0828	1.1505	1.6178	0.068	0.641	0.641
	$\boldsymbol{\lambda}$	0.0022	0.0052	0.0045	0.0020	0.00095	0.0005
1000	$\boldsymbol{\mu}$	0.0048	0.7055	1.2222		0.180	0.180
	$\boldsymbol{\sigma}$	0.00521	0.16407	0.27833	0.00507	0.06179	0.0618

Table 1: Estimated Mean Squared Errors from 300 replications of plain Gaussian EM and constrained EC-MM algorithms started from the true parameters, for the parallel test synthetic model, and two sample sizes.

here. These inversions only occurred for the $n = 100$ case.

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

Table 2: Estimated Mean Squared Errors from 300 replications of plain Gaussian EM and constrained EC-MM algorithms for the tau equivalent synthetic model, and two sample sizes.

5 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. Scientifically, however, we had a reason—from the psychometric example motivating this article—to place the constraints on the parameter space in any case. For this reason, we do not consider the question here of any type of hypothesis test of a null hypothesis such as "the constraints hold", and we compare estimates only on the basis of their mean squared errors rather than, say, BIC or AIC. However, testing and model selection methods generally require estimation algorithms for their implementation, and the algorithms we present could certainly serve this purpose if future work were to address these additional issues.

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 algorithms presented in this article are implemented in the R package called mixtools (Benaglia et al., 2009) for the R statistical software (R Development Core Team, 2010), which is publicly available on the Comprehensive R Archive Network (CRAN) at `cran.r-project.org`.

References

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.

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.

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

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

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

Kim, D. and Taylord, J. (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.

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. (1992). The EM algorithm for latent class analysis with equality constraints. *Psychometrika*, 2:261–269.

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

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

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.

Shi, N., Zheng, S., 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.

Young, D. S., Benaglia, T., Chauveau, D., Elmore, R. T., Hettmansperger, T. P., Hunter, D. R., Thomas, H., and Xuan, F. (2009). mixtools: Tools for mixture models. R package version 0.3.3.