Biometrika (2011), **98**, 2, *pp.* 403–416
doi: 10.1093/biomet/asq079
Advance Access publication 28 April 2011

Maximum smoothed likelihood for multivariate mixtures

BY M. LEVINE

Department of Statistics, Purdue University, West Lafayette, Indiana 47907, U.S.A.
mlevins@purdue.edu

D. R. HUNTER

Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16801, U.S.A
dhunter@stat.psu.edu

AND D. CHAUVEAU

Laboratoire de Mathématiques - Analyse, Probabilités, Modélisation - Orléans, Université d'Orléans, 45067 Orléans cedex 2, France
didier.chauveau@univ-orleans.fr

SUMMARY

We introduce an algorithm for estimating the parameters in a finite mixture of completely unspecified multivariate components in at least three dimensions under the assumption of conditionally independent coordinate dimensions. We prove that this algorithm, based on a majorization-minimization idea, possesses a desirable descent property just as any EM algorithm does. We discuss the similarities between our algorithm and a related one, the so-called non-linearly smoothed EM algorithm for the non-mixture setting. We also demonstrate via simulation studies that the new algorithm gives very similar results to another algorithm that has been shown empirically to be effective but that does not satisfy any descent property. We provide code for implementing the new algorithm in a publicly available R package.

Some key words: EM algorithm; Majorization-minimization algorithm; Nonlinearly smoothed EM algorithm; Nonparametric mixture.

1. INTRODUCTION

Suppose the r-dimensional vectors $X_1, \ldots, X_n$ are a simple random sample from a finite mixture density of m components $f_1, \ldots, f_m$, with $m > 1$ and known in advance. It is assumed throughout this paper that each of these densities f_j equals the product of its marginal densities:

$$f_j(x) = \prod_{k=1}^{r} f_{jk}(x_k). \tag{1}$$

Taking a fully nonparametric approach with regard to estimating the f_{jk}, we may therefore express the finite mixture density as

$$X_i \sim g_\theta(x_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jk}(x_{ik}), \tag{2}$$

where $\lambda = (\lambda_1, \ldots, \lambda_m)$ must satisfy

$$\sum_{j=1}^{m} \lambda_j = 1, \quad \lambda_j \geqslant 0. \tag{3}$$

Here, we assume $X_i = (X_{i1}, \ldots, X_{ir})^{\mathrm{T}}$ and we let θ denote the vector of parameters to be estimated, including the mixing proportions $\lambda_1, \ldots, \lambda_m$ and the univariate densities f_{jk}. Throughout, j and k always denote the component and coordinate indices, respectively.

According to (2), conditional on knowing the particular subpopulation the observation X_j came from, its coordinates are independent. This conditional independence assumption has appeared in a growing body of literature on non- and semi-parametric multivariate mixture models; see, for example, Benaglia et al. (2009a). Conditional independence may be thought of as a simplification of the commonly used repeated measures random effects model, which often assumes that multivariate observations on an individual are independent, conditional on the identity of the individual in question. Here, we simply replace the individual-level effects of a repeated measures model by component-level effects. We demonstrate the usefulness of this model on real data in § 5·3.

The question of identifiability of the parameters in (2) is of central theoretical importance. By identifiability, we refer to the question of when g_θ uniquely determines λ and each of the f_{jk}, at least up to label-switching and any changes to the densities f_{jk} that occur on a set of Lebesgue measure zero that therefore do not change the distributions F_{jk}. Here, label-switching refers to permuting the order of the summands $1, \ldots, m$ in (2). Hall & Zhou (2003) established that when $m = 2$, identifiability of parameters generally follows in $r \geqslant 3$ dimensions but not fewer. However, a general result for more than two components proved elusive, though several articles on the topic have appeared (e.g., Hall et al., 2005; Kasahara & Shimotsu, 2008). Finally, Allman et al. (2009) proved an elegant and powerful result using a theorem of Kruskal (1977), establishing the identifiability of the parameters in (2) whenever $r \geqslant 3$, regardless of m, as long as the density functions $f_{1k}, \ldots, f_{mk}$ are linearly independent except possibly on a set of Lebesgue measure zero.

An EM-like algorithm designed to estimate θ in (2) was introduced in Benaglia et al. (2009a). That algorithm is much simpler than alternatives in the literature and yields considerably smaller mean integrated squared errors than an alternative algorithm (Hall et al., 2005) in a simulation study. However, the algorithm lacks any sort of theoretical justification and can only be called EM-like because it resembles an EM algorithm in its formulation. The current paper corrects this shortcoming by introducing a smoothed loglikelihood function and formulating a new iterative algorithm with a provable monotonicity property that produces results similar to those of Benaglia et al. (2009a) in tests.

The association of EM algorithms with mixture models has a long history: the paper in which the initials EM were coined (Dempster et al., 1977) describes a finite mixture model as one of several examples. A sizable literature on nonparametric mixtures appeared (see Lindsay, 1995, Ch. 5–8) in which the mixing distribution is nonparametric, unlike the current paper in which the component distributions are nonparametric. For the earlier notion of nonparametric mixture models, Vardi et al. (1985) introduced an EM algorithm for maximum likelihood estimation of the mixing distribution that has an elegant convergence theory but that unfortunately does not deal with ill-posedness of the problem. Silverman et al. (1990) remedied this shortcoming with the smoothed EM algorithm that smoothes the result of each step of the classical EM algorithm. The practical performance of this algorithm is excellent but it is difficult to analyse theoretically. In an unpublished 1992 paper, Eggermont first proposed the idea of using a regularization approach

to modify the smoothed EM algorithm in such a way that the resulting algorithm is easier to investigate theoretically. Eggermont & LaRiccia (1995) showed that the resulting nonlinear smoothed EM algorithm is a true EM algorithm and that its convergence theory is very similar to that of the EM algorithm introduced in the mixing density estimation context by Shepp & Vardi (1982). The nonlinear smoothed EM may be shown to have a unique solution and its practical performance is competitive with alternatives (Eggermont, 1999).

The current paper unites the two different meanings of nonparametric mixture by introducing an algorithm inspired by the nonlinear smoothed EM that does in fact possess a descent property and that converges to a local maximizer of a likelihood-like quantity for the finite mixture model (2) in which the component densities are completely unspecified. The likelihood-like function used is similar to that introduced in Eggermont & LaRiccia (1995) for an unspecified continuous mixing distribution; it is, essentially, a penalized Kullback–Leibler distance between the target density function and the iteratively reweighted sum of smoothed component density function estimates. For its optimization, we rely on a computational tool called a majorization-minimization algorithm, which is yet another generalization of an EM algorithm (Hunter & Lange, 2004).

2. Smoothing the log-density

Let us assume that Ω is a compact subset of $\mathbb{R}^r$ and define the linear vector function space

$$\mathcal{F} = \{f = (f_1, \ldots, f_m)^{\mathrm{T}} : 0 < f_j \in L_1(\Omega), \log f_j \in L_1(\Omega), j = 1, \ldots, m\}.$$

The assumption of compact support may appear limiting, but it is not problematic from a practical point of view and plays no role in the implementation of the algorithm we propose.

Take $K(\cdot)$ to denote some kernel density function on the real line. With a slight abuse of notation, let us define the product kernel function $K(u) = \prod_{k=1}^r K(u_k)$ and its rescaled version $K_h(u) = h^{-r} \prod_{k=1}^r K(h^{-1}u_k)$. Furthermore, we define a smoothing operator $\mathcal{S}$ for any function $f \in L_1(\Omega)$ by

$$\mathcal{S}f(x) = \int_\Omega K_h(x-u) f(u)\, du.$$

Furthermore, we extend $\mathcal{S}$ to $\mathcal{F}$ by defining $\mathcal{S}f = (\mathcal{S}f_1, \ldots, \mathcal{S}f_m)^{\mathrm{T}}$. We also define a nonlinear smoothing operator $\mathcal{N}$ as

$$\mathcal{N}f(x) = \exp\{(\mathcal{S}\log f)(x)\} = \exp \int_\Omega K_h(x-u) \log f(u) du.$$

This operator is strictly concave, and it is also multiplicative in the sense that $\mathcal{N}f_j = \prod_k \mathcal{N}f_{jk}$ for f_j defined as in (1). The concavity is proved as Lemma 3.1(iii) of Eggermont (1999). The idea of smoothing the logarithm of the density function goes back to Silverman (1982), where a penalty based on the second derivative of the log-density is discussed.

To simplify notation, we introduce the finite mixture operator

$$\mathcal{M}_\lambda f(x) = \sum_{j=1}^m \lambda_j f_j(x),$$

whence we also obtain $\mathcal{M}_\lambda f(x) = g_\theta(x)$ and

$$\mathcal{M}_\lambda \mathcal{N} f(x) = \sum_{j=1}^m \lambda_j \mathcal{N} f_j(x).$$

Let $g(x)$ now represent a known target density function. We begin by defining the following functional of θ and, implicitly, g:

$$\ell(\theta) = \int_{\Omega} g(x) \log \frac{g(x)}{(\mathcal{M}_\lambda \mathcal{N} f)(x)}\, dx. \tag{4}$$

We will suppress the subscripted Ω on the integral sign from now on. Our goal in § 3 will be to find a minimizer of $\ell(\theta)$ subject to the assumptions that each f_{jk} is a univariate density function and λ satisfies (3).

Remark 1. An immediate consequence of (4) is that $\ell(\theta)$ can be viewed as a penalized Kullback–Leibler distance between $g(x)$ and $(\mathcal{M}_\lambda N f)(x)$. Indeed, if we define

$$D(a \mid b) = \int \left\{ a(x) \log \frac{a(x)}{b(x)} + b(x) - a(x) \right\} dx$$

as usual, it follows that

$$\ell(\theta) = D(g \mid \mathcal{M}_\lambda \mathcal{N} f) + \int g(x) dx - \sum_{j=1}^{m} \lambda_j \int \mathcal{N} f_j(x) dx,$$

where $-\lambda_j \int \mathcal{N} f_j(x)\, dx$ is a penalization term; see Eggermont 1999, equation (1.12) and the discussion immediately following.

3. A MAJORIZATION-MINIMIZATION ALGORITHM

Our goal is to define an iterative algorithm that possesses a descent property with respect to the functional $\ell(f, \lambda)$; that is, we wish to ensure that the value of $\ell(f, \lambda)$ cannot increase from one iteration to the next. We write $\ell(f, \lambda)$ instead of $\ell(\theta)$ so we may discuss f and λ separately. Suppose that we were to define an iteration operator G, to be applied to the vector $f = (f_1, \ldots, f_m)^{\mathrm{T}}$, as

$$G f_j(x) = \alpha_j \int K_h(x - u) \frac{g(u) \mathcal{N} f_j(u)}{\mathcal{M}_\lambda \mathcal{N} f(u)}\, du$$

for each j, where α_j is a proportionality constant that makes $G f_j(\cdot)$ integrate to one. Using an argument analogous to the proof of Lemma 1, we could show that

$$\ell(f, \lambda) - \ell(Gf, \lambda) \geqslant \sum_{j=1}^{m} \frac{\lambda_j}{\alpha_j} \int G f_j(x) \log \frac{G f_j(x)}{f_j(x)}\, dx = \sum_{j=1}^{m} \frac{\lambda_j}{\alpha_j} D(G f_j | f_j) \geqslant 0.$$

Thus, the above definition evidently results in an algorithm that satisfies the descent property. Unfortunately, however, it does not preserve the essential conditional independence assumption (1). We must therefore use a slightly different approach.

Let (f^0, λ^0) denote the current parameter values in an iterative algorithm. Our strategy for minimizing $\ell(f, \lambda)$ is based on a majorization-minimization algorithm, in which we define a functional $b^0(f, \lambda)$ that, when shifted by a constant, majorizes $\ell(f, \lambda)$. That is,

$$b^0(f, \lambda) + C^0 \geqslant \ell(f, \lambda), \quad \text{with equality when } (f, \lambda) = (f^0, \lambda^0). \tag{5}$$

The superscript on b^0 indicates that the definition of $b^0(f, \lambda)$ will in general depend on the parameter values (f^0, λ^0), which are considered fixed constants in this context.

Majorization-minimization algorithms are not solely applicable to the current context, but enjoy widespread usage in statistical problems in which optimization of a difficult objective function may be avoided by iteratively optimizing a series of simpler functions. A general introduction to these algorithms, which are sometimes more appropriately called minorization-maximization algorithms using the same initials, is given by Hunter & Lange (2004).

For $j = 1, \ldots, m$, let

$$w_j^0(x) = \frac{\lambda_j^0 \mathcal{N} f_j^0(x)}{\mathcal{M}_{\lambda^0} \mathcal{N} f^0(x)}. \tag{6}$$

The weight functions w_j^0 satisfy $\sum_j w_j^0(x) = 1$. We now claim that

$$b^0(f, \lambda) = -\int g(x) \sum_{j=1}^m w_j^0(x) \log\{\lambda_j \mathcal{N} f_j(x)\}\, dx \tag{7}$$

gives the majorizing functional we seek.

LEMMA 1. *The function of* (7) *satisfies* $\ell(f, \lambda) - \ell(f^0, \lambda^0) \leqslant b^0(f, \lambda) - b^0(f^0, \lambda^0)$.

Proof. By the convexity of the logarithm function and the fact that $\sum_j w_j^0(x) = 1$,

$$\begin{aligned}
\ell(f, \lambda) - \ell(f^0, \lambda^0) &= -\int g(x) \log \frac{\sum_{j=1}^m \lambda_j \mathcal{N} f_j(x)}{\mathcal{M}_{\lambda^0} \mathcal{N} f^0(x)}\, dx \\
&= -\int g(x) \log \sum_{j=1}^m w_j^0(x) \frac{\lambda_j \mathcal{N} f_j(x)}{\lambda_j^0 \mathcal{N} f_j^0(x)}\, dx \\
&\leqslant -\int g(x) \sum_{j=1}^m w_j^0(x) \log \frac{\lambda_j \mathcal{N} f_j(x)}{\lambda_j^0 \mathcal{N} f_j^0(x)}\, dx \\
&= b^0(f, \lambda) - b^0(f^0, \lambda^0). \qquad \square
\end{aligned}$$

Lemma 1 verifies the majorization claim (5), where we take the constant C^0 to be $\ell(f^0, \lambda^0) - b^0(f^0, \lambda^0)$.

Rewriting (7), we obtain

$$b^0(f, \lambda) = -\sum_{j=1}^m \sum_{k=1}^r \iint K_h(x_k - u) g(x) w_j^0(x) \log f_{jk}(u)\, du\, dx - \sum_{j=1}^m \log \lambda_j \int g(x) w_j^0(x)\, dx. \tag{8}$$

Above, and henceforth, u denotes a scalar, whereas x is an r-dimensional vector. Also, $b^0(f, \lambda)$ separates the parameters from each other, in the sense that it is the sum of separate functions of the individual f_{jk} and λ_j.

Subject to the constraint $\sum_j \lambda_j = 1$, it is not hard to minimize $b^0(f, \lambda)$ with respect to the λ parameter: for each j, the minimizer is

$$\hat{\lambda}_j = \frac{\int g(x) w_j^0(x)\, dx}{\sum_{j=1}^m \int g(x) w_j^0(x)\, dx} = \int g(x) w_j^0(x)\, dx. \tag{9}$$

Next, let us focus on only the part of (8) involving f_{jk} by defining

$$b^0_{jk}(f_{jk}) = -\iint K_h(x_k - u)g(x)w^0_j(x)\log f_{jk}(t)\,du\,dx.$$

LEMMA 2. *For $j = 1, \ldots, m$ and $k = 1, \ldots, r$, define*

$$\hat{f}_{jk}(u) = \alpha_{jk}\int K_h(x_k - u)g(x)w^0_j(x)\,dx, \tag{10}$$

where α_{jk} is a constant chosen so that $\int \hat{f}_{jk}(u)\,dt = 1$. Then $\hat{f}_{jk}$ is the unique, up to changes on a set of Lebesgue measure zero, density function minimizing $b^0_{jk}(\cdot)$.

Proof. Fubini's theorem yields

$$b^0_{jk}(f_{jk}) = \frac{1}{\alpha_{jk}}D(\hat{f}_{jk} \mid f_{jk}) - \frac{1}{\alpha_{jk}}\int \hat{f}_{jk}(u)\log \hat{f}_{jk}(u)\,du,$$

where the second term on the right-hand side does not depend on f_{jk}. The result follows immediately. □

Let us now combine the preceding results. From Lemma 1, we conclude that

$$\ell(\hat{f}, \hat{\lambda}) - \ell(f^0, \lambda^0) \leqslant b^0(\hat{f}, \hat{\lambda}) - b^0(f^0, \lambda^0). \tag{11}$$

Furthermore, we know from Lemma 2 and (9) that each individual piece of the $b^0(\cdot)$ function of (8) is minimized by the corresponding piece of $(\hat{f}, \hat{\lambda})$. We conclude that the right-hand side of inequality (11) is bounded above by zero, which proves the descent property summarized by the following theorem.

THEOREM 1. *Define $\hat{\lambda}$ as in (9) and $\hat{f}$ as in Lemma 2. Then $\ell(\hat{f}, \hat{\lambda}) \leqslant \ell(f^0, \lambda^0)$.*

4. ESTIMATING THE PARAMETERS

We now assume that we are given a simple random sample $x_1, \ldots, x_n$ distributed according to the $g_\theta(x)$ density defined in (2). Letting $\tilde{G}_n(\cdot)$ denote the empirical distribution function of the sample and ignoring the term $\int g_\theta(x)\log g_\theta(x)\,dx$ whose empirical version does not involve any parameters, a discrete version of (4) is

$$\ell_n(f, \lambda) = \int \log \frac{1}{(\mathcal{M}_\lambda \mathcal{N} f)(x)}\,d\tilde{G}_n(x) = -\sum_{i=1}^n \log\{(\mathcal{M}_\lambda \mathcal{N} f)(x_i)\}.$$

Here, $\ell_n(f, \lambda)$ resembles a penalized loglikelihood function except for the presence of the nonlinear smoothing operator $\mathcal{N}$ and the fact that with the negative sign preceding the sum, our goal is minimization rather than maximization of $\ell_n(\cdot)$.

Using an argument nearly identical to the one leading to (10), we may show that the following algorithm ensures that the value of $\ell_n(\cdot)$ decreases at each iteration.

Given initial values (f^0, λ^0), iterate the following three steps for $t = 0, 1, \ldots$:

(i) Majorization step. Define, for each i and j,

$$w^t_{ij} = \frac{\lambda^t_j \mathcal{N} f^t_j(x_i)}{\mathcal{M}_{\lambda^t}\mathcal{N} f^t(x_i)} = \frac{\lambda^t_j \mathcal{N} f^t_j(x_i)}{\sum_{a=1}^m \lambda_a \mathcal{N} f^t_a(x_i)}. \tag{12}$$

(ii) Minimization step, part 1. Set

$$\lambda_j^{t+1} = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^t \tag{13}$$

for $j = 1, \ldots, m$.

(iii) Minimization step, part 2. For each j and k, let

$$f_{jk}^{t+1}(u) = \frac{\sum_{i=1}^n w_{ij}^t K_h(u - x_{ik})}{\sum_{i=1}^n w_{ij}^t} = \frac{1}{nh\lambda_j^{t+1}} \sum_{i=1}^{n} w_{ij}^t K\left(\frac{u - x_{ik}}{h}\right). \tag{14}$$

Equations (12), (13) and (14) are merely the discrete versions of (6), (9) and (10), respectively. With regard to the convergence properties of the algorithm we have defined here, we prove in the Appendix that, if we hold λ fixed and repeatedly iterate (10), then the sequence of f functions converges to a global minimizer of $\ell(f, \lambda)$ for that value of λ.

5. Implementation and numerical examples

5·1. *Blocks of identically distributed coordinates*

As in Benaglia et al. (2009a), we can extend the model of conditional independence to a more general model: we allow the existence of blocks of coordinates that are identically distributed. If we let b_k denote the block index of the kth coordinate, where $1 \leqslant b_k \leqslant B$ and B is the total number of such blocks, then (2) is replaced by

$$g_\theta(x_i) = \sum_{j=1}^{m} \lambda_j \prod_{k=1}^{r} f_{jb_k}(x_{ik}). \tag{15}$$

These blocks may all be of size one as in (2) if $b_k = k$ $(k = 1, \ldots, r)$, or there may exist only a single block if $b_k = B = 1$ $(k = 1, \ldots, r)$, which is the case of conditionally independent and identically distributed variables. The nonlinear smoothing operator $\mathcal{N}$ applied to f_j is simply $\mathcal{N} f_j = \prod_{k=1}^r \mathcal{N} f_{jb_k}$, and definitions of $\mathcal{M}_\lambda f$ and $\mathcal{M}_\lambda \mathcal{N} f$ are unchanged.

The algorithm of § 4 can easily be adapted for handling the block structure. In fact, both the majorization step (12) and the first part of the minimization step (13) remain unchanged. The second part of the minimization step (14) becomes:

(iv) Minimization step, part 2, modified. For each component j and block $\ell \in \{1, \ldots, B\}$, let

$$\begin{aligned} f_{j\ell}^{t+1}(u) &= \frac{\sum_{k=1}^r \sum_{i=1}^n w_{ij}^t I_{\{b_k=\ell\}} K_h\,(u - x_{ik})}{\sum_{k=1}^r \sum_{i=1}^n w_{ij}^t I_{\{b_k=\ell\}}} \\ &= \frac{1}{nh\lambda_j^{t+1} C_\ell} \sum_{k=1}^{r} \sum_{i=1}^{n} w_{ij}^t I_{\{b_k=\ell\}} K\left(\frac{u - x_{ik}}{h}\right), \end{aligned}$$

where $C_\ell = \sum_{k=1}^r I_{\{b_k=\ell\}}$ is the number of coordinates in the ℓth block.

This algorithm is implemented as the npMSL function in the latest version, currently version 0·4·4, of the publicly available R package (R Development Core Team, 2011) called mixtools (Young et al., 2010; Benaglia et al., 2009b). The block structure imposes a type of constraint on the density functions in that certain subsets of these functions are assumed to be the same. Further constraints, such as densities that are assumed to have the same shape but differ only in a location

or scale parameter, are discussed by Benaglia et al. (2011); in principle, such constraints should be straightforward to incorporate using our algorithm.

5·2. *Simulated examples*

Our first simulation study compares the nonparametric EM-like algorithm from Benaglia et al. (2009a) with the new algorithm using the same examples for which Hall et al. (2005) tested their estimation technique based on inverting the mixture model. The three simulated models, described below, are trivariate two-component mixtures ($m = 2, r = 3$) with independent but not identically distributed repeated measures, i.e. $b_k = k$ ($k = 1, 2, 3$). We ran $S = 300$ replications of $n = 500$ observations each and computed the errors in terms of the square root of the mean integrated squared error for the densities, where

$$\text{MISE}_{jk} = \frac{1}{S}\sum_{s=1}^{S}\int \{\hat{f}_{jk}^{(s)}(u) - f_{jk}(u)\}^2\,du, \quad j = 1, 2\ k = 1, 2, 3;$$

and the integral is computed numerically using an appropriate function defined in mixtools. Each density $\hat{f}_{jk}^{(s)}$ is computed using the weighted kernel density estimate (14) together with the final values of the posterior probabilities p_{ij}^t after convergence of the algorithm.

The first example is a normal model, for which the individual densities $f_{j\ell}$ are $\mathcal{N}(\mu_{j\ell}, 1)$, with component means $\mu_1 = (0, 0, 0)$ and $\mu_2 = (3, 4, 5)$. The second example uses double exponential distributions with densities $f_{j\ell}(t) = \exp(-|t - \mu_{j\ell}|)/2$, where $\mu_1 = (0, 0, 0)$ and $\mu_2 = (3, 3, 3)$. In the third example, the first component has a central $t(10)$ distribution and thus $\mu_1 = (0, 0, 0)$, whereas the second component's coordinates are noncentral $t(10)$ distributions with noncentrality parameters 3, 4, and 5. Thus, the mean of the third component is $\mu_2 = (3, 4, 5) \times 1{\cdot}0837$. Both algorithms assume only the general model of conditional independence, with $b_k = k$ for all k.

Since Benaglia et al. (2009a) showed that the nonparametric EM dramatically outperforms the inversion method of Hall et al. (2005) for the three test cases, Fig. 1 only compares the nonparametric EM against the new algorithm, which is labelled smoothed in the figure. This figure shows that the two algorithms provide nearly identical efficiency in terms of mean integrated squared error and that there is no clear winner for the models and the various settings considered.

A second numerical study compares the smoothed algorithm with a parametric Gaussian EM algorithm for model (15) both when the true model is Gaussian and when it has heavier tails. The simulated data have $m = 2$ components and are trivariate with conditionally independent and identically distributed coordinates. A parametric Gaussian EM algorithm for this situation is implemented in the mixtools package as the function repnormmixEM(). We ran the simulations for a Gaussian mixture of $\mathcal{N}(0, 1)$ and $\mathcal{N}(3, 1)$, and a mixture of $t(5)$ and $t'(5, 3)$ so that the mean of the noncentral t distribution is 3·57, with $\lambda_1 = 0{\cdot}3$ in both cases. Table 1 shows comparisons based on the component weight estimate $\hat{\lambda}_1$ and mean estimates $\hat{\mu}_1$ and $\hat{\mu}_2$, though these values are not estimated directly in the nonparametric case and must be computed as functions of the other parameter estimates. Our smoothed algorithm performs as well as the parametric Gaussian EM algorithm when the model is correctly specified and, as expected, far better than the parametric EM algorithm when the model is slightly misspecified.

Finally, we tested our algorithm for $r \leqslant 2$, where identifiability may fail. Our experience suggests that parameter estimates are often sensible when $r = 2$ but nonsensical when $r = 1$. The reason for the $r = 2$ success may be the almost identifiability (Hall & Zhou, 2003) that holds in this case, yet further research appears necessary to determine when it is reasonable to proceed

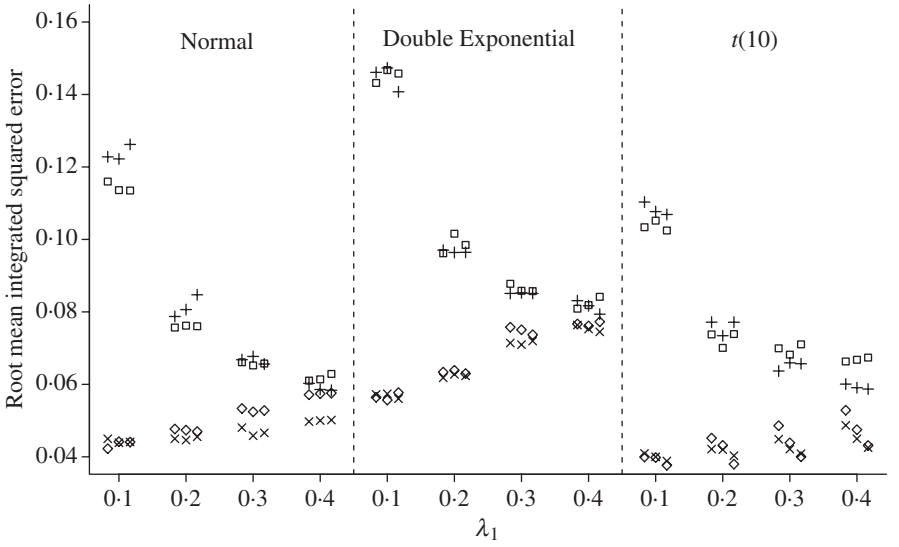

Fig. 1. Root mean integrated squared error of estimated densities for $\lambda_1 \in \{0{\cdot}1, 0{\cdot}2, 0{\cdot}3, 0{\cdot}4\}$ for the three benchmark models from Hall et al. (2005). Estimates for all three coordinates are shown, with horizontal values jittered to improve visibility. □, smoothed component 1; ◇, smoothed component 2; +, nonparametric EM component 1; and ×, nonparametric EM component 2.

Table 1. *Estimated bias, standard deviation and mean squared error for the Gaussian* EM *and nonparametric algorithms based on* 300 *replications for each model with sample size* $n = 500$. *True parameter values of* λ_1, μ_1, *and* μ_2 *are* 0·3, 0, *and* 3 *or* 3·57, *respectively*

		Gaussian EM			Nonparametric smoothed		
		λ_1	μ_1	μ_2	λ_1	μ_1	μ_2
Gaussian	Mean	0·3002	0·0028	2·9997	0·3001	0·0033	2·9994
	SD	0·0195	0·0501	0·0323	0·0195	0·0501	0·0322
	MSE	0·00038	0·00252	0·00104	0·00038	0·00252	0·00104
$t(5)$	Mean	0·36	0·77	3·15	0·299	0·008	3·568
	SD	0·18	1·56	1·07	0·0207	0·0689	0·0586
	MSE	0·036	3·012	1·319	0·00043	0·00482	0·00344

with estimation when $r = 2$. Certainly, estimation in the univariate $r = 1$ case is not viable without additional restrictions such as those of Bordes et al. (2006) and Hunter et al. (2007).

5·3. *The water-level experiment*

We consider in this section a dataset from an experiment involving $n = 405$ children aged 11 to 16 years subjected to a water-level task as initially described by Thomas et al. (1993). In this experiment, each child is presented with eight rectangular vessels on a sheet of paper, each tilted to one of $r = 8$ clock-hour orientations: in order of presentation to the subjects, these orientations are 11, 4, 2, 7, 10, 5, 1 and 8 o'clock. The children's task was to draw a line representing the surface of still liquid in the closed, tilted vessel in each picture. Each such line describes two points of intersection with the sides of the vessel; the acute angle, in degrees, formed between the horizontal and the line passing through these two points was measured for each response. The sign of each such measurement was taken to be the sign of the slope of the line. The water-level dataset is available in the mixtools package (Young et al., 2010; Benaglia et al., 2009b). This dataset has

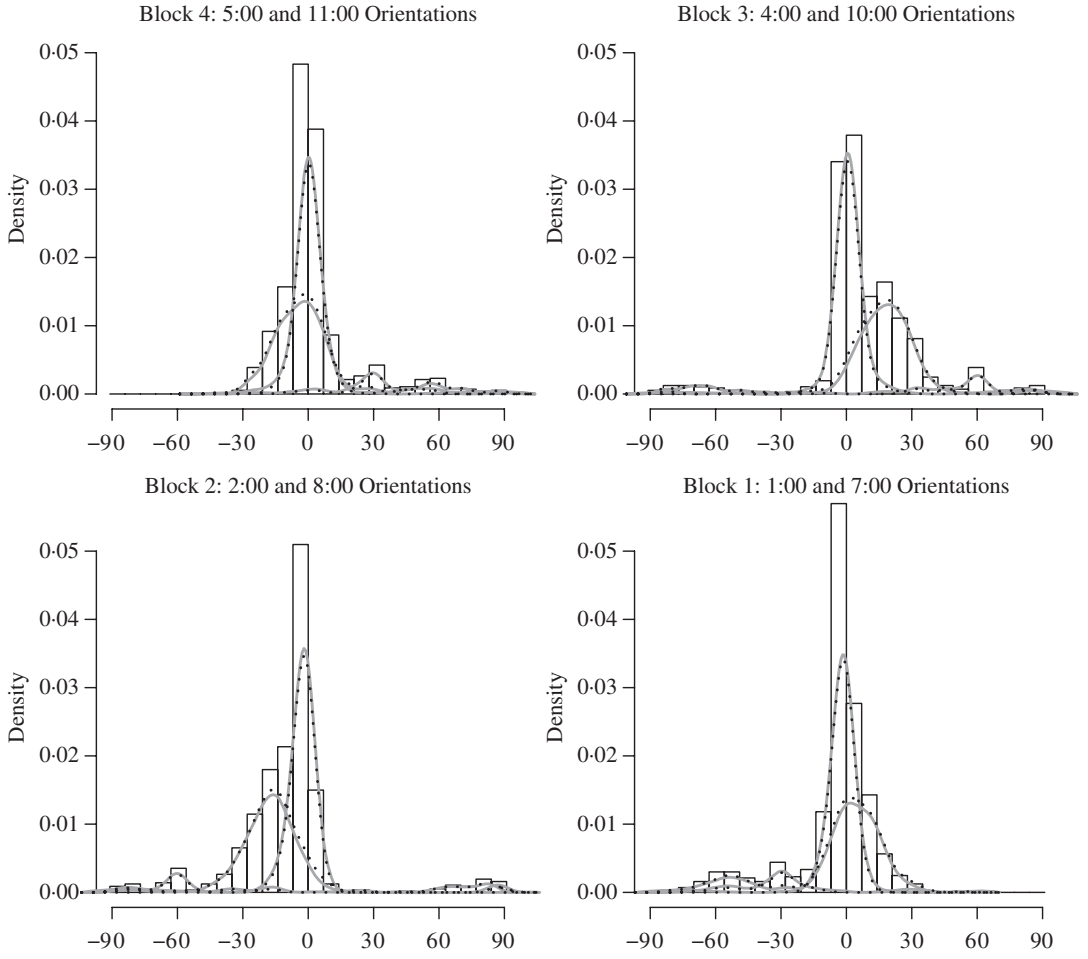

Fig. 2. The water-level data, analysed using the Benaglia et al. (2009a) algorithm (solid grey lines) and the new smoothed algorithm (dotted lines), assuming (15) with $m = 3$ mixture components and bandwidth $h = 4$.

been analysed previously by Hettmansperger & Thomas (2000) and Elmore et al. (2004), who assume that the $r = 8$ coordinates are all conditionally identically distributed and then bin the data to produce multinomial vectors. These authors call this a cut-point approach.

However, because of the experimental methodology used to collect the data, it seems reasonable to weaken the assumption that each orientation's measurements are identically distributed; instead, we only assume that opposite clock-face orientations lead to conditionally independent and identically distributed responses, so that the eight coordinates may be organized into four blocks of two each, where the densities within each block are identical, which is (15).

Benaglia et al. (2009a) apply their algorithm to (15) with $B = 4$ and blocks of coordinates defined by $b = (b_1, \ldots, b_8) = (4, 3, 2, 1, 3, 4, 1, 2)$, which means, e.g. that $b_4 = b_7 = 1$, i.e. block 1 relates to coordinates 4 and 7, corresponding to clock orientations 1:00 and 7:00.

Figure 2 compares, for $m = 3$ components, the nonparametric EM solution with the solution given by the new algorithm of § 5·1, which we refer to as the smoothed algorithm. Although the overlapping lines in the figure make it difficult to see exactly which component captures which mode, it is clear that the two solutions are overall quite similar: each appears to detect one component of children who understand the task, the component peaked around the correct answer of zero degrees, another group who appear to complete the task correctly when the vessel is near

Table 2. *Estimates and* 95% *confidence intervals for* λ*, based on* 10 000 *bootstrap replications, for the water-level data example*

	λ_1	λ_2	λ_3
Nonparametric EM	0·492 (0·446, 0·552)	0·431 (0·337, 0·469)	0·077 (0·054, 0·151)
Smoothed	0·471 (0·420, 0·527)	0·465 (0·361, 0·496)	0·064 (0·052, 0·159)

vertical but who do not do as well in the sideways orientations of blocks 2 and 3, and a third group who appear to draw the line perpendicular to the sides of the vessel. The estimated proportions of these three components for the smoothed algorithm, along with the corresponding nonparametric EM estimates, are summarized in Table 2. The confidence intervals in the table were computed using a nonparametric bootstrap approach by repeatedly resampling with replacement from the empirical distribution defined by the n observed r-dimensional vectors and carefully checking for label-switching occurrences in the resulting estimates. Here, label-switching refers to permuting the three labels on $\hat{\lambda}_1$, $\hat{\lambda}_2$ and $\hat{\lambda}_3$, which in this example is easy to detect by examining standard deviations of the estimated densities in combination with the $\hat{\lambda}$ estimates obtained. We see that our bootstrapped confidence intervals are not centred at the estimates in all cases, though this is not unusual in our experience.

The two algorithms took a comparable number of iterations to converge, 69 on average for the smoothed algorithm versus 71 for the nonparametric EM algorithm, using the same convergence criterion, though each iteration of the smoothed algorithm took roughly 1/3 longer due to the numerical convolution involved. Overall, despite the small differences in the estimates obtained by the two algorithms, we do not notice a systematic pattern in these differences and the results are really quite close.

6. Discussion

The algorithm we propose in this paper is a refinement of the algorithm proposed in Benaglia et al. (2009a), which, despite its many favourable qualities, does not minimize any particular objective function. In contrast, the algorithm we introduce has a provable descent property with respect to a smoothed loglikelihood while obtaining parameter estimates similar to the original algorithm. The price paid for the theoretically desirable descent property appears to be a small loss of computational speed, since our algorithm involved numerical convolutions.

The majorization technique used in our majorization-minimization algorithm, exploiting the convexity of the negative logarithm in the proof of Lemma 1, is the same technique as is used by a classical EM algorithm. However, the question of whether our algorithm represents a true EM algorithm is of little importance; our paper demonstrates how a direct majorization approach can produce the same theoretical advantages as an EM approach. There are some corresponding disadvantages, such as a slow linear rate of convergence as the algorithm nears an optimum point. As a potential improvement, one might consider acceleration techniques (Lange et al., 2000).

The basic model of (2) may be generalized in several directions. In addition to the blocking structure introduced in (15), various location and/or scale models are possible. A thorough discussion of such generalizations is given in § 4 of Benaglia et al. (2009a). Our algorithm may easily be adapted to these situations.

Unlike the implicitly defined solution attained by the Benaglia et al. (2009a) algorithm, our algorithm achieves a local maximum of an explicit smoothed loglikelihood function. This should facilitate theoretical development and possibly lead to inferential techniques that may be used in conjunction with this algorithm. Empirical studies in Benaglia et al. (2009a) already suggest approximate rates of convergence of estimates to the true values in simulation studies as n grows.

Since our algorithm appears to produce very similar estimates, presumably its convergence rates are also similar; the theory in this paper should promote work on this conjecture.

Sensible choice of the bandwidth h is a challenging problem. We have simplified the discussion by assuming a common h for each component and coordinate, or block, though it is straightforward to introduce component- and block-specific bandwidths $h_{j\ell}$. Complicating the selection of the bandwidth in the mixture setting is the fact that one does not observe individual component densities. This issue is discussed at length by Benaglia et al. (2011), who recommend iteratively updating the bandwidths to exploit the knowledge about component assignments as it is gained. Doing this for our algorithm appears to destroy the desirable descent property. Thus, effective automatic bandwidth selection remains a topic for further inquiry.

ACKNOWLEDGEMENT

The work of Michael Levine is partially supported by a grant from the Division of Mathematical Sciences of the National Science Foundation. The authors thank the reviewers and associate editor for their thorough reading and helpful comments.

APPENDIX

Some convergence properties

Fix λ^0 and consider the function defined by (10) that maps $(f^0, \lambda^0) \mapsto (\hat{f}, \lambda^0)$. Iteratively applying this function yields a sequence

$$(f^0, \lambda^0), (f^1, \lambda^0), (f^2, \lambda^0), \ldots . \tag{A1}$$

Here, we present a few simple convergence results regarding this sequence. The results of this section have analogues in § 3 of Eggermont (1999) for a slightly different case.

We assume that the kernel $K(\cdot)$ is strictly positive on the whole real line. Define $B \subset \mathcal{F}$ by

$$B = \left\{ \mathcal{S}\phi : 0 \leqslant \phi \in \mathcal{F}, \int_\Omega \phi_j(x)\, dx = 1, j = 1, \ldots, m \right\}.$$

The idea is that B will contain the whole sequence $f^0, f^1, f^2, \ldots$ except possibly the initial f^0, where each element in the sequence is defined by applying (10) to the preceding element. To verify this claim, observe that (10) may be rewritten as

$$\hat{f}_{jk}(u) = \mathcal{S}\phi^0_{jk}(u), \tag{A2}$$

where

$$\phi^0_{jk}(x_k) = \alpha_{jk} \int \cdots \int g(x) w^0_j(x)\, dx_1 \cdots dx_{k-1} dx_{k+1} \cdots dx_r \tag{A3}$$

must integrate to one because of the definition of α_{jk}. Furthermore, the functional $f \mapsto \ell(f, \lambda)$ is defined on B because for any $(f_1, \ldots, f_m) \in B$, f_j is bounded below by $\inf_{x\in\Omega} K_h(x)$, which is positive since Ω is compact and K is positive. Thus, $\mathcal{N}f$ is well-defined for $f \in B$.

LEMMA A1. *The set B is convex and $\ell(f, \lambda)$ of (4) is strictly convex on B for fixed λ.*

Proof. For $f^1, f^2 \in B$ and $\alpha \in (0, 1)$, the linearity of $\mathcal{S}$ implies $\alpha f^1 + (1-\alpha) f^2 \in B$. Also,

$$\ell\{\alpha f^1 + (1-\alpha) f^2\} = \int g(x) \log g(x)\, dx - \int g(x) \log \sum_{j=1}^m \lambda_j \mathcal{N}\{\alpha f^1_j + (1-\alpha) f^2_j\}(x)\, dx.$$

Focusing on the rightmost term above, we first claim that

$$\mathcal{N}\{\alpha f^1_j + (1-\alpha) f^2_j\}(x) > \alpha \mathcal{N} f^1(x) + (1-\alpha) \mathcal{N} f^2(x)$$

by the strict concavity of the $\mathcal{N}$ operator (Lemma 3.1(iii) of Eggermont, 1999). Furthermore, the fact that the logarithm function is concave and strictly increasing implies that

$$\begin{aligned}\ell\{\alpha f^1 + (1-\alpha) f^2, \lambda\} &< \int g(x) \log g(x)\, dx - \alpha \int g(x) \log \sum_{j=1}^{m} \lambda_j (\mathcal{N} f_j^1)(x)\, dx \\ &\quad - (1-\alpha) \int g(x) \log \sum_{j=1}^{m} \lambda_j (\mathcal{N} f_j^2)(x)\, dx \\ &= \alpha \ell(f^1, \lambda) + (1-\alpha) \ell(f^2, \lambda). \end{aligned}$$

□

Remark 1. Using nearly the same proof as for Lemma A1, we may show that $\ell(f, \lambda)$ is strictly convex in λ for fixed f. However, no strict convexity is possible in general for $\ell(f, \lambda)$ since, as in all mixture model settings, permuting the subscripts $1, \ldots, m$ on $(f_1, \lambda_1), \ldots, (f_m, \lambda_m)$ does not change the value of $\ell(f, \lambda)$; thus, there exists no unique global minimizer of $\ell(f, \lambda)$.

The following lemma adds the assumption of Lipschitz continuity to those mentioned earlier to establish a sufficient condition guaranteeing that the sequence of functions in (A1) has a uniformly convergent subsequence.

LEMMA A2. *If there exists $L > 0$ such that $|K_h(x) - K_h(y)| \leqslant L|x - y|$ for any $x, y \in \Omega$, then every functional sequence $f^1, f^2, \ldots$ defined by* (A1) *has a uniformly convergent subsequence.*

Proof. Since Ω is a compact subset of $\mathbb{R}^r$, we may assume that there exist positive constants $a < A$ such that $a \leqslant K_h(\cdot) \leqslant A$ on Ω. Thus, in (A2) and (A3), we must have $a \leqslant \hat{f}_{jk} \leqslant A$ for all j, k. We conclude that the sequence $|f^1|, |f^2|, \ldots$ is uniformly bounded.

Furthermore, for arbitrary $x, y \in \Omega$ and $f \in B$,

$$\begin{aligned}|f_j(x) - f_j(y)| &= |S\phi_j(x) - S\phi_j(y)| \\ &\leqslant \int |K_h(x-u) - K_h(y-u)||\phi_j(u)|\, du \leqslant L|x-y| \end{aligned}$$

for all j. We conclude that the sequence $f^1, f^2, \ldots$ is uniformly bounded and equicontinuous, so the Arzelà–Ascoli theorem implies that there is a uniformly convergent subsequence. □

LEMMA A3. *The functional $f \mapsto \ell(f, \lambda)$ is lower semicontinuous on B.*

Proof. Consider a sequence of functions $\{f_n\} = \{(f_{1,1}, \ldots, f_{m,n})^{\mathrm{T}}\} \in B$. Let us denote $\psi = (\psi_1, \ldots, \psi_m)^{\mathrm{T}} = \liminf_n f_n$. By Lemma A2, there always exists a subsequence $f_{n_k} \to \psi$; without loss of generality, assume that this subsequence coincides with the entire sequence $\{f_n\}$. Since every component function $f_{j,n} \in B$ is bounded away from zero, so is the limit function ψ; therefore, $\log f_{j,n} \to \log \psi$. Consequently, $\mathcal{N} f_{j,n} \to \mathcal{N}\psi_j$ and $\mathcal{M}_\lambda \mathcal{N} f_n \to \mathcal{M}_\lambda N \psi$. Since the function $\rho(t) = t - \log t - 1 \geqslant 0$, Fatou's lemma gives

$$\int g(x) \rho\{M_\lambda N \psi(x)\}\, dx \leqslant \liminf \int g(x) \rho\{M_\lambda N f_n(x)\}\, dx.$$

From the above, the statement of the proposition follows immediately. □

The functional $f \mapsto \ell(f, \lambda)$ is uniformly bounded from below on B, which follows from (5) and the fact that $\mathcal{N} f_j(x) \leqslant \mathcal{S} f_j(x)$ by the arithmetic-geometric mean inequality. Thus, the lower semicontinuity combined with strict convexity, as proved above, imply that for any fixed λ, sequence (A1) converges to a global maximizer of the functional $f \mapsto \ell(f, \lambda)$. As a practical matter, this means that we could essentially

replace $\ell(f, \lambda)$ by the profile loglikelihood

$$\ell^*(\lambda) = \inf_{f \in B} \ell(f, \lambda)$$

because the minimization on the right-hand side may be accomplished by iterating (10) until convergence. However, dealing with the profile loglikelihood is not the general optimization strategy adopted in § 4.

REFERENCES

ALLMAN, E. S., MATIAS, C. & RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37**, 3099–132.

BENAGLIA, T., CHAUVEAU, D. & HUNTER, D. R. (2009a). An EM-like algorithm for semi-and non-parametric estimation in multivariate mixtures. *J. Comp. Graph. Statist.* **18**, 505–26.

BENAGLIA, T., CHAUVEAU, D. & HUNTER, D. R. (2011). Bandwidth selection in an EM-like algorithm for nonparametric multivariate mixtures. In *Nonparametrics and Mixture Models: A Festschrift Dedicated to Thomas P. Hettmansperger*, Ed. D. R. Hunter, D. S. P. Richards and J. L. Rosenberger. Singapore: World Scientific.

BENAGLIA, T., CHAUVEAU, D., HUNTER, D. R. & YOUNG, D. (2009b). mixtools: An R package for analyzing finite mixture models. *J. Statist. Software* **32**, 1–29.

BORDES, L., MOTTELET, S. & VANDEKERKHOVE, P. (2006). Semiparametric estimation of a two-component mixture model. *Ann. Statist.* **34**, 1204–32.

DEMPSTER, A. P., LAIRD, N. M. & RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc.* B **39**, 1–38.

EGGERMONT, P. P. B. (1999). Nonlinear smoothing and the EM algorithm for positive integral equations of the first kind. *Appl. Math. Optim.* **39**, 75–91.

EGGERMONT, P. P. B. & LARICCIA, V. N. (1995). Maximum smoothed density estimation for inverse problems. *Ann. Statist.* **23**, 199–220.

ELMORE, R. T., HETTMANSPERGER, T. P. & THOMAS, H. (2004). Estimating component cumulative distribution functions in finite mixture models. *Commun. Statist. Theory Meth.* **33**, 2075–86.

HALL, P., NEEMAN, A., PAKYARI, R. & ELMORE, R. T. (2005). Nonparametric inference in multivariate mixtures. *Biometrika* **92**, 667–78.

HALL, P. & ZHOU, X. H. (2003). Nonparametric estimation of component distributions in a multivariate mixture. *Ann. Statist.* **31**, 201–24.

HETTMANSPERGER, T. P. & THOMAS, H. (2000). Almost nonparametric inference for repeated measures in mixture models. *J. R. Statist. Soc.* B **62**, 811–25.

HUNTER, D. R. & LANGE, K. (2004). A tutorial on MM algorithms. *Am. Statistician* **58**, 30–7.

HUNTER, D. R., WANG, S. & HETTMANSPERGER, T. P. (2007). Inference for mixtures of symmetric distributions. *Ann. Statist.* **35**, 224–51.

KASAHARA, H. & SHIMOTSU, K. (2008). Nonparametric identification of finite mixture models of dynamic discrete choices. *Econometrica* **77**, 135–76.

KRUSKAL, J. B. (1977). Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Lin. Algeb. Applic.* **18**, 95–138.

LANGE, K., HUNTER, D. R. & YANG, I. (2000). Optimization transfer using surrogate objective functions. *J. Comp. Graph. Statist.* **9**, 1–20.

LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry, and Applications*. Institute of Mathematical Statistics.

R DEVELOPMENT CORE TEAM (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

SHEPP, L. A. & VARDI, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imag.* **1**, 113–22.

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795–810.

SILVERMAN, B. W., JONES, M. C., WILSON, J. D. & NYCHKA, D. W. (1990). A smoothed EM algorithm approach to indirect estimation problems, with particular reference to stereology and emission tomography. *J. R. Statist. Soc.* B **52**, 271–324.

THOMAS, H., LOHAUS, A. & BRAINERD, C. J. (1993). Modeling growth and individual differences in spatial tasks. *Monog. Soc. Res. Child Dev.* **58**, 1–191.

VARDI, Y., SHEPP, L. A. & KAUFMAN, L. (1985). A statistical model for positron emission tomography. *J. Am. Statist. Assoc.* **80**, 8–20.

YOUNG, D. S., BENAGLIA, T., CHAUVEAU, D., ELMORE, R. T., HETTMANSPERGER, T. P., HUNTER, D. R., THOMAS, H. & XUAN, F. (2010). Mixtools: tools for analyzing finite mixture models. R package version 0.4.4.

[*Received June* 2010. *Revised November* 2010]