Biometrika (2001), **88**, 2, *pp.* 535–550

Printed in Great Britain

Alternative EM methods for nonparametric finite mixture models

BY RAMANI S. PILLA

Division of Epidemiology & Biostatistics, 2121 West Taylor Street, University of Illinois, Chicago, Illinois 60612, U.S.A.

pillar@uic.edu

AND BRUCE G. LINDSAY

Department of Statistics, The Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.

bgl@psu.edu

SUMMARY

This research focuses on a general class of maximum likelihood problems in which it is desired to maximise a nonparametric mixture likelihood with finitely many known component densities over the set of unknown weight parameters. Convergence of the conventional EM algorithm for this problem is extremely slow when the component densities are poorly separated and when the maximum likelihood estimator requires some of the weights to be zero, as the algorithm can never reach such a boundary point. Alternative methods based on the principles of EM are developed using a two-stage approach. First, a new data augmentation scheme provides improved convergence rates in certain parameter directions. Secondly, two 'cyclic versions' of this data augmentation are created by changing the missing data formulation between the EM-steps; these extend the acceleration directions to the whole parameter space, giving another order of magnitude increase in convergence rate. Examples indicate that the new cyclic versions of the data augmentation schemes can converge up to 500 times faster than the conventional EM algorithm for fitting nonparametric finite mixture models.

Some key words: Augmentation; Complete data; EM algorithm; Finite mixture distribution; High-dimensional; Maximum likelihood; Missing data; Rate of convergence; Nonparametric mixture; Zero-elimination.

1. INTRODUCTION AND MOTIVATION

Consider a class of statistical problems that require maximisation of a mixture likelihood with finitely many known component densities over the set of unknown weight parameters, here denoted by π's; we call the corresponding model the 'fixed support mixture model'. The maximisation problem arises in several statistical contexts, including mixture maximum likelihood estimation with known component densities (Lindsay, 1995, Ch. 1), optimal design (Silvey, 1980), interval censored data (Böhning et al., 1996) and indirectly observed Poisson processes including the Positron Emission Tomography problem (Vardi et al., 1985). Roeder et al. (1989) provide an extensive study of the theoretical properties of maximum likelihood in this context, as well as applications in genetics.

In certain cases, one can apply the methods developed here to problems in which the component densities have unknown parameters. To fix ideas, consider the class of binomial mixtures, in which each observed y has, conditional on θ, a Bi(n, θ) distribution, where the latent parameter θ arises from an unknown mixing distribution $G(\theta)$. One could fix the number of components in the mixing distribution, but allow the θ parameters to be unknown for each component. We will call this the 'continuous support mixture model'. One can fit this model by using a continuous support EM algorithm (Dempster et al., 1977) which estimates the θ parameters simultaneously with the π parameters. Although the methods developed here could be extended in various ways to this model, we do not consider them in this paper.

In other cases, one may be ignorant of the appropriate number of binomial components and wish to maximise the likelihood over all the mixing distributions G, giving the 'nonparametric maximum likelihood estimator' which is a discrete distribution on the parameter space with a random number of component densities (Lindsay, 1983). We call this the 'nonparametric mixture model'. One can determine the nonparametric maximum likelihood estimator via a fixed support mixture algorithm as follows: (i) construct a fine grid of θ values and consider them to be the support set of G; then (ii) maximise over the π parameters on this support, leaving θ values fixed. Once one has diagnosed the number of active components through the nonzero π values, further refinements can be made on the θ parameters using a secondary algorithm; see Böhning et al. (1992) and Böhning et al. (1998) for details.

The examples in § 6 will be drawn from this second class of problems because they can be used to investigate the roles of many overlapping densities in a straightforward way by increasing the number and closeness of the points on the grid; this creates an ideal situation in terms of mimicking a large-scale problem. In addition, the choice of examples will enable us to compare the grid approach with approaches that would use a continuous support EM in finding the nonparametric maximum likelihood estimator; see § 6·3. However, the main emphasis of this research is to improve upon the existing fixed support mixture EM method.

In the fixed support mixture case, the convergence of the EM algorithm is extremely slow when the component densities are similar and when the maximum likelihood solution requires some of the weight parameters to be zero, as the algorithm can never reach such a boundary point. This paper develops and investigates EM-based methods to address this problem.

We refer to McLachlan & Krishnan (1997) and the references therein for several extensions of EM and for a number of comparisons between different approaches existing in the literature. Celeux et al. (1996) compared three different stochastic versions of the EM algorithm, namely stochastic EM (SEM), simulated annealing EM (SAEM) and Monte Carlo EM (MCEM), on the mixture problem through intensive Monte Carlo numerical simulations and a real-data study. They showed that, for some mixtures, SEM is almost always preferable to EM, SAEM and MCEM. They also mentioned that, for severely overlapping mixtures, none of these algorithms can be confidently used. Jamshidian & Jennrich (1997) proposed new quasi-Newton-based EM accelerators which accelerated EM by factors of over 100 in some cases. However, their mixture examples only dealt with two-component mixtures of unknown component densities. Neal & Hinton (1999) show empirically that their incremental variant of the EM gives faster convergence in two-component mixtures of unknown normal densities.

Meng & van Dyk (1997) considered speeding up convergence of EM by 'efficient' ways

of augmenting data; by 'efficient', they mean less augmentation of observed data while maintaining the stability and simplicity of EM. While one of our goals here is to augment with fewer missing data, we focus on some new and very specific augmentation schemes and cyclical approaches. A particular feature of our strategy is to stay within the formal EM framework, because of its simplicity and reliability in very high-dimensional problems. The approach relies on the following two basic principles.

(1) *Reducing the complete data.* A key point in the mixture case is that there are various ways to specify the 'missing data' and so there is no unique EM algorithm. That is, the form of the algorithm depends on the specification of the 'complete data', which consists of observed data and hypothetical missing data.

(2) *Cycling between data augmentations.* As EM progresses through the parameter space, it has least favourable directions in which the convergence is inherently slow. These directions depend, in turn, on the augmented data structure. The second main idea is to construct a sequence of complementary missing data structures such that, by cycling through them, one achieves improved acceleration in all directions.

The remainder of the paper is organised as follows. We set up the mixture problem in § 2. In §§ 3 and 4 we present a new data augmentation scheme and its cyclical approaches, respectively. Section 5 discusses the empirical issues. The various approaches are compared using two examples in § 6. Section 7 provides possible extensions and some discussion.

2. The conventional EM algorithm in the mixture problem

Consider the following optimisation problem: given parameters $\pi = (\pi_1, \ldots, \pi_m)$ in the unit simplex, maximise the nth degree polynomial in π given by $L(\pi) = \prod_i \sum_j \pi_j f_j(y_i)$, where $f_j(y_i) \geqslant 0$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$, are known constants. In particular, $L(\pi)$ is commonly the likelihood function for a random sample $y = (y_1, \ldots, y_n)$ from a mixture of known component densities $f_j(y)$, where π is a vector of unknown mixing weights.

We will present our analyses here in terms of a multinomial likelihood, presenting the necessary modifications for the continuous case at the end of the section. Consider an independently and identically distributed sample from a discrete mixture density function

$$\mathrm{pr}(Y = t \mid \pi) = g(t \mid \pi) := \sum_{j=1}^{m} \pi_j f_j(t),$$

with sample space $\{0, \ldots, T\}$. In this case, the observed likelihood of a sample $y_1, \ldots, y_n$ becomes

$$L_{\mathrm{obs}}(\pi \mid n) = \prod_{t=0}^{T} \{g(t \mid \pi)\}^{n(t)},$$

where $n = \{n(0), \ldots, n(T)\}$ are the observed data.

In the multinomial case, one can regard the observed data $n(t)$ as the column marginal totals of an unobserved table of counts, $\{n_j(t)\}$, where $n_j(t)$, in row j and column t, is the unobserved number of observations from component j that were equal to t. This leads to the conventional missing data formulation in which the complete data vector is $x = (n_1, \ldots, n_m)$, where $n_j = \{n_j(0), \ldots, n_j(T)\}$ for all $j = 1, \ldots, m$. In this scenario, the likelihood for the complete data becomes

$$L_{\mathrm{com}}(\pi \mid x) := \prod_{t=0}^{T} \prod_{j=1}^{m} \{\pi_j f_j(t)\}^{n_j(t)}.$$

The EM algorithm based on this complete data likelihood will be called the 'conventional

EM' algorithm sometimes shortened to EM algorithm; see Lindsay (1995, pp. 62–3) for further details.

Remark 1. The change in the EM formulae for the continuous case merely involves setting the range of t to be the observed y_i and letting $n(t) = 1$.

Remark 2. This problem is a special case of the nonparametric mixture maximum likelihood problem. Among other features, the maximum likelihood estimate $\hat{\pi}$ is unique in the sense described by Lindsay (1995, p. 60).

Remark 3. When we say 'conventional EM', we are referring to the conventional EM for the fixed support mixture problem in which the component densities are fixed and known, and we wish to find the global maximum over all such mixtures.

3. Reducing the complete data

3·1. *Preamble*

Our goal is to find alternative complete data structures with 'less missing data' than conventional EM, which is at its worst in determining the relative weights $\pi_j/(\pi_j + \pi_k)$ and $\pi_k/(\pi_j + \pi_k)$ given to two densities $f_j(t)$ and $f_k(t)$ that are very similar.

One way to motivate the problem of similar densities is to view the mixture model as a linear model with constraints on the π_j parameters. Let $g'_\pi = \{g(0|\pi), \ldots, g(T|\pi)\}$. We can write the vector of probabilities g_π as a linear combination of component density vectors f_j: $g_\pi = \pi_1 f_1 + \ldots + \pi_m f_m = F\pi$. When the component densities are not well separated, the columns in F are highly correlated. As in linear regression with multicollinearity, this means that the parameter values can be varied widely with little effect on the density, creating in turn a relatively flat likelihood function. This multicollinearity also has an adverse effect on the EM algorithm, as then the missing data become much more informative relative to the observed data.

Our methods will take account of the most highly correlated pairs, since the component densities are known, to reduce the effect of multicollinearity.

3·2. *Partially missing data structure*

We assume that m, the total number of densities, is even. Our new data augmentation method starts by pairing the densities, $\{(f_1, f_2), \ldots, (f_{m-1}, f_m)\}$, say, such that the pairing reflects correlated pairs; the last density f_m is treated separately if m is odd. We consider the consequences of having the 'reduced complete data'

$$\{N_1(.) = n_1(.) + n_2(.), \ldots, N_{m/2}(.) = n_{m-1}(.) + n_m(.)\};$$

recall that the complete data for EM is the contingency table $\{n_j(.)\}$. Since the dimension of the missing data is reduced from $(m-1)$ to $(m/2)-1$, about half, one would expect a corresponding improvement in performance. However, this reduction creates a new problem in that there is no longer an explicit solution to the M-step, clear from the case $m = 2$, where the augmented data equals the observed data.

We now show that nonetheless the augmented data $\mathcal{N}^* = \{N_1, \ldots, N_{m/2}\}$, where $N_k = \{N_k(0), \ldots, N_k(T)\}$ for all $k = 1, \ldots, m/2$, generates a simple optimisation problem even in higher dimensions. We call $\mathcal{N}^*$ the 'paired complete data' and the corresponding EM algorithm the 'paired complete data EM algorithm', sometimes shortened to the paired EM.

In this scenario, the complete data likelihood function becomes

$$L_{\mathrm{PCD}}(\pi \mid \mathscr{N}^*) \propto \prod_{t=0}^{T} \{\pi_1 f_1(t) + \pi_2 f_2(t)\}^{N_1(t)} \ldots \{\pi_{m-1} f_{m-1}(t) + \pi_m f_m(t)\}^{N_{m/2}(t)}.$$

Reparameterisation using $\alpha_j = \pi_j/(\pi_j + \pi_{j+1})$ and $\alpha_{j+1} = 1 - \alpha_j$, for all odd j, yields

$$L_{\mathrm{PCD}}(\pi^*, \alpha \mid \mathscr{N}^*) \propto \prod_{t=0}^{T} (\pi_1 + \pi_2)^{N_1(t)} \{\alpha_1 f_1(t) + \alpha_2 f_2(t)\}^{N_1(t)} \ldots$$
$$(\pi_{m-1} + \pi_m)^{N_{m/2}(t)} \{\alpha_{m-1} f_{m-1}(t) + \alpha_m f_m(t)\}^{N_{m/2}(t)}, \qquad (3{\cdot}1)$$

where $\pi^* = (\pi_1^* = \pi_1 + \pi_2, \ldots, \pi_{m/2}^* = \pi_{m-1} + \pi_m)$ and $\alpha = (\alpha_1, \ldots, \alpha_m)$ are the new parameters to be estimated. Note that $\pi_1^* + \ldots + \pi_{m/2}^* = 1$ and $\alpha_j + \alpha_{j+1} = 1$ for all odd $j \in \{1, \ldots, m\}$, so we still have $(m-1)$ unknowns, and that the α parameters are the relative weights of the pairs.

The function $\log L_{\mathrm{PCD}}(\pi^*, \alpha \mid \mathscr{N}^*)$ can be written in the form

$$\mathscr{A}(\pi^*) + \sum_{\text{odd } j} \mathscr{B}(\alpha_j),$$

where

$$\mathscr{A}(\pi^*) = \sum_t \sum_k N_k(t) \log \pi_k^*, \quad \mathscr{B}(\alpha_j) = \sum_t N_k(t) \log\{\alpha_j f_j(t) + (1 - \alpha_j) f_{j+1}(t)\}.$$

Given the parameter values π^{*p} and α^p at stage p, the $(p+1)$st step of the corresponding EM algorithm involves maximisation of

$$Q(\pi^*, \alpha \mid \pi^{*p}, \alpha^p) := E\{\log L_{\mathrm{PCD}}(\pi^*, \alpha \mid \mathscr{N}^*) \mid n;\ \pi^{*p}, \alpha^p\}.$$

The E-step replaces $N_k(t)$ in $\mathscr{A}(\pi^*)$ and $\mathscr{B}(\alpha_j)$ with their conditional expectations, given the observed data, n, and the current parameter values (π^{*p}, α^p), namely

$$\tilde{N}_k^p(t) = n(t) \pi_k^{*p} \frac{\{\alpha_j^p f_j(t) + (1 - \alpha_j^p) f_{j+1}(t)\}}{g(t \mid \pi^{*p}, \alpha^p)}, \qquad (3{\cdot}2)$$

where $g(t \mid \pi^{*p}, \alpha^p) = \sum_{\text{odd } j} \pi_k^{*p} \{\alpha_j^p f_j(t) + (1 - \alpha_j^p) f_{j+1}(t)\}$. The optimisation problem in the M-step is simplified because of the separated parameters. We can maximise $\mathscr{A}(\pi^*)$ explicitly, subject to the constraint $\sum \pi_k^* = 1$. The $\mathscr{B}(\alpha_j)$ terms do not generate explicit solutions, but they are univariate functions that can be optimised using reliable one-dimensional methods such as Newton–Raphson. Applying a single Newton–Raphson step on each $\mathscr{B}(\alpha_j)$ in the M-step turns out to be highly effective, as Newton–Raphson for the univariate parameter α_j is not only quadratically convergent but also is nearly monotonic (Böhning & Lindsay, 1988) and so is highly reliable here.

Lange (1995, Proposition 1) has shown that the rate of convergence of EM is independent of the number of Newton–Raphson steps within each EM-step, so one Newton–Raphson step is sufficient, and is our recommendation. The modifications proposed by Böhning & Lindsay (1988) to adjust Newton–Raphson for guaranteed monotonicity has proven to be unnecessary in practice and hence was not implemented in the examples considered in § 6. Note that we do not maximise the expected loglikelihood in the M-step, but rather take a step towards maximisation.

3·3. *Practical implementation of paired EM*

Safeguards. We implemented a number of safeguards to avoid an indeterminate form in the intermediate stages of the algorithm.

(i) If both π_j and π_{j+1} are zeros simultaneously, then $\alpha_j = \pi_j/\pi_{j/2}$ becomes an indeterminate form, so we set $\alpha_j = \frac{1}{2}$.

(ii) If the first and second derivatives of $\mathscr{B}(\alpha_j)$ with respect to α_j, $\dot{\mathscr{B}}(\alpha_j)$ and $\ddot{\mathscr{B}}(\alpha_j)$, are both zeros, the Newton–Raphson step is skipped; this is the case when the filled-in missing counts $\tilde{N}_k(t)$ given by equation (3·2) are zeros or the absolute difference between the paired densities, $|f_j(t) - f_{j+1}(t)|$, is too close to zero over the range of the data.

Parameter bounds. The following technique modifies the Newton–Raphson step so that the updated α_j satisfies $0 \leqslant \alpha_j \leqslant 1$. If Newton–Raphson over-steps, resulting in $\alpha_j > 1$, we check whether or not $\dot{\mathscr{B}}(\alpha_j) < 0$ at $\alpha_j = 1$. If so, then the global maximum is to the left of 1, and we proceed back towards that solution by taking one Newton–Raphson step from $\alpha_j = 1$. The resulting solution definitely lies in the interval $[0, 1]$ (Böhning & Lindsay, 1988). If $\dot{\mathscr{B}}(\alpha_j) > 0$ at $\alpha_j = 1$, then the unconstrained solution is to the right of 1, which implies that the global constrained solution sets α_j to 1. If Newton–Raphson steps too far left, resulting in $\alpha_j < 0$, we follow the same procedure with a reversal of all inequalities.

Logistic transformation. An alternative formulation is to reparameterise the α_j parameters in the Newton–Raphson step using the logistic transformation, $\gamma_j = \log\{\alpha_j/(1-\alpha_j)\}$, thereby eliminating the need for the preceding boundary checks. However, this can adversely affect the rate of convergence when the solution is zero or 1, as Lange's result applies only when the solution in the γ parameters is in the parameter set $(-\infty, \infty)$. Note that in the α-parameterisation one can set α to zero in a single step, whereas the γ can never reach zero. Our empirical results, unpublished, showed that the convergence rate in the γ-parameterisation depended on the number of Newton–Raphson steps used within each EM-step, and the rate was inferior to that of the α-parameterisations.

Delta EM. Our concerns about the reliability and monotonicity of the Newton–Raphson step in paired EM led us to develop another algorithm which we called delta EM (Pilla & Lindsay, 1996). The delta EM algorithm gives explicit solutions to the M-step, but it is inferior to paired EM both in theoretical efficiency and in numerical performance, so was excluded from this presentation; see Pilla & Lindsay (1996) or R. S. Pilla's (1997) Ph.D. dissertation from the Department of Statistics at the Pennsylvania State University for details.

3·4. *The elimination of zero support points*

The fact that paired EM gains speed from pairing means that its performance might be improved by sequentially eliminating from the algorithm those component densities that have mass zero. We illustrate through a simple case of four densities. Suppose that $f_1, \ldots, f_4$ are the original densities, with corresponding weights $\pi_1, \ldots, \pi_4$. Suppose that in the pairs (f_1, f_2) and (f_3, f_4) the weights π_2 and π_3 are set to zero. To make the algorithm more efficient, we eliminate the f_2 and f_3 densities from the original pairs and consider the new pair (f_1, f_4). As we will see in § 6, when carefully done this 'zero-elimination' can greatly improve the convergence rate of the algorithms when many parameters are approaching zero. This procedure can be fully automated, as will be described in § 5·3.

4. TWO NEW COMPLETE DATA CYCLES

4·1. *Introduction*

The paired complete data augmentation scheme depends on a selected pairing of densities $\{(f_1, f_2), (f_3, f_4), \ldots\}$. Paired EM shows acceleration in finding the relative weights, α,

within each pair, while providing no major improvement in the total weights, π^*, yielding overall only 2 to 9 times acceleration in terms of computational speed, as will be shown in § 6. That is, as these algorithms progress through the parameter space, they have least favourable directions, corresponding to π^*, in which the convergence is inherently slow. These directions depend, in turn, on the pairing in the augmentation. We next consider changing the missing data formulation between the EM-steps to achieve improved acceleration in all directions.

4·2. *Rotation cycles*

The first and simplest cyclic approach involves rotating through a sequence of different pairings of densities. Within each pairing scheme, we use one step of paired EM, and then move to the next pairing scheme. The pairing schemes are constructed such that the paired densities are similar, so that paired EM is most effective. We will show that the acceleration effects are spread throughout the parameter space when the various pairings are chosen to be complementary; if one does not change the pairing, there is only a small improvement over conventional EM.

We illustrate this with eight densities, where the adjacent pairs are most similar. In this case, one could alternate between pairing (f_1, f_2), (f_3, f_4), (f_5, f_6) and (f_7, f_8) in an odd EM-step and (f_2, f_3), (f_4, f_5), (f_6, f_7) and (f_8, f_1) in an even EM-step. Each step then accelerates a complementary set of relative weights. As we will see, this very simple two-step cycle is highly effective, particularly when the solution $\hat{\pi}$ has many zeros. Intuitively, this is because the transferral of mass between the densities can flow unimpeded as one proceeds through a series of cycles. For example, mass can efficiently shift from f_1 to f_2 in one step, from f_2 to f_3 in the next step, and so forth, thereby facilitating the movement of mass towards any one density that will receive a large final weight.

When the paired complete data augmentation scheme is implemented with the above two-step rotation cycles, we shall refer to it simply as 'rotated EM'.

4·3. *Hierarchical cycles*

The second cyclic approach is based upon the observation that one can use the paired complete data augmentation scheme to accelerate the determination of the relative weights of any set of pairings of blocks of densities. We illustrate this 'paired block EM' principle first, before turning it into a cyclic approach.

Consider the m-component model with component indices $i = 1, \ldots, m$ and break the indices into B, assuming it is even, blocks,

$$A_1 = \{1, \ldots, a_1\}, \quad A_2 = \{a_1 + 1, \ldots, a_2\}, \quad \ldots, \quad A_{B-1} = \{a_{B-2} + 1, \ldots, a_{B-1}\},$$

$$A_B = \{a_{B-1} + 1, \ldots, a_B \equiv m\},$$

say. The mixture density can now be rewritten as $g(t \mid \pi) = \sum \pi_i^\dagger P_i(t)$, where $\pi_i^\dagger = \sum_{j \in A_i} \pi_j$ for all $i = 1, \ldots, B$ and the 'relative block densities' are $P_i(t) = \sum_{j \in A_i} \alpha_j^\dagger f_j(t)$ with $\alpha_j^\dagger = \pi_j / \pi_i^\dagger$, where $\pi_i^\dagger = \sum_{j \in A_i} \pi_j$. We then apply paired EM to the densities $P_i(t)$, optimising over the $\pi_i^\dagger$ parameters while holding the $\alpha_j^\dagger$ parameters fixed to accelerate the relative weights $\pi_j^\dagger / (\pi_j^\dagger + \pi_{j+1}^\dagger)$ of adjoining block densities. To do so, we reparameterise the π

parameters as

$$\pi_k^* = \sum_{j \in A_{2k-1}} \pi_j + \sum_{j \in A_{2k}} \pi_j = \pi_{2k-1}^\dagger + \pi_{2k}^\dagger, \tag{4·1}$$

$$\alpha_{2k-1} = \frac{\pi_{2k-1}^\dagger}{\pi_{2k-1}^\dagger + \pi_{2k}^\dagger} = \frac{\pi_{2k-1}^\dagger}{\pi_k^*} \tag{4·2}$$

and $\alpha_{2k} = 1 - \alpha_{2k-1}$, for all $k = 1, \ldots, B/2$. Thus $\pi^\dagger$ is the sum of all the π parameters in a given block and π^* is the sum of the two $\pi^\dagger$ parameters of paired blocks. See the website http://www.uic.edu/~pillar/ for paired block EM.

4·4. *The hierarchical paired complete data EM algorithm*

We illustrate with eight densities how to construct hierarchical parameterisation and turn it into a sequential EM.

Step 1. Apply one step of paired EM to the first set of pairs, $(f_1, f_2), \ldots, (f_7, f_8)$, to find the relative weights, $\alpha_1 = \pi_1/\pi_1^*, \ldots, \alpha_7 = \pi_7/\pi_4^*$, and total weights,

$$\pi_1^* = \pi_1 + \pi_2, \quad \ldots, \quad \pi_4^* = \pi_7 + \pi_8.$$

This accelerates the relative weights α but not the π^* parameters; to achieve this, we move up to another level in the hierarchy.

Step 2. Apply block EM on the blocks $A_1 = (1, 2), \ldots, A_4 = (7, 8)$, where indices in each block identify the component densities to be grouped. That is, pair blocks (A_1, A_2) and (A_3, A_4) and find the relative block densities, P_i, and π^* parameters as in § 4·3 to accelerate the relative weights

$$\beta_1 = \frac{\pi_1 + \pi_2}{\pi_1 + \ldots + \pi_4} = \frac{\pi_1^\dagger}{\pi_1^*}, \quad \beta_2 = \frac{\pi_5 + \pi_6}{\pi_5 + \ldots + \pi_8} = \frac{\pi_2^\dagger}{\pi_2^*}.$$

The β parameters correspond to the α parameters given by equation (4·2). This level accelerates β parameters but not the π^* parameters; to achieve this, we move up to another level.

Step 3. Merge each pair of blocks from Step 2 into a single block to form new blocks $A_1 = (1, 2, 3, 4)$ and $A_2 = (5, 6, 7, 8)$. By redefining the P_i densities and π^* parameters, once again apply one step of paired EM to estimate the relative weight $\pi_1^\dagger/\pi_1^* = (\pi_1 + \ldots + \pi_4)$. Note that $\pi_1^* = 1$. This is actually a conditional maximisation step, as a single pair has no missing data.

Henceforth, EM based on the paired complete data with hierarchical EM cycles will be referred to as 'hierarchical EM'. Moreover, one can create an even larger cyclic scheme by shifting the density pairings between each application of the hierarchical cycle. If done by rotation, henceforth this will be referred to as 'composite EM'.

Remark 4. The above hierarchical scheme is easily extended to any number of densities by taking a sequence of block EM's such that each block EM in the cycle accelerates estimates of relative weights, α, within pairs of blocks, whereas the block totals, the π^* parameters, are not enhanced until the next level of hierarchy; see the Appendix mentioned below for details. The scheme is most easily implemented when there are 2^R densities, for any positive integer R, in which case there are R EM-steps within each cycle; for details

for the case when the number of densities, m, is not a power of 2, see the Appendix of this paper available at http://www.uic.edu/~pillar/.

Remark 5. As one might expect, it is not necessary to cycle through the blocking schemes in the order given above. One could, for example, also go in reverse order, starting with two large paired blocks and then subdividing, sequentially. It was shown in R. S. Pilla's Ph.D. dissertation that the asymptotic rate of convergence of a hierarchical cycle is the same for both the methods. One could also choose the block EM's in a random order; we doubt that the possible numerical benefits would be worth the programming effort.

5. Empirical issues

5·1. *Empirical assessment of rate of convergence*

In comparing various algorithms, our focus will be on

$$\Lambda^p = \{\log L_{\text{obs}}(\hat{\pi}\,|\,n) - \log L_{\text{obs}}(\pi^p\,|\,n)\},$$

which we call the 'residual' of the loglikelihood at the pth iteration. The most important measurement of the convergence of a maximum likelihood algorithm is the value of the loglikelihood, as it provides information about the accuracy of the parameter estimates on a confidence interval scale (Lindsay, 1995, pp. 131–2). It follows that a natural summary of the speed of a linearly convergent algorithm like EM is the rate of convergence of the observed loglikelihood to its maximum value. The following lemma shows that, for a linearly convergent algorithm, the plot of $\log_{10}\Lambda^p$ against p should become linear with slope equal to $\log_{10}\mathrm{r}$, where r is the asymptotic rate of convergence of the sequence $\{\log L_{\text{obs}}(\pi^p\,|\,n)\}_{p=0,1,\ldots}$ at $\log L_{\text{obs}}(\hat{\pi}\,|\,n)$ generated by any of the several EM algorithms. Thus the smaller the r for any given loglikelihood sequence, the faster it is progressing towards the maximum likelihood estimate.

Lemma 1. *If the sequence* $\{\log L_{\text{obs}}(\pi^p\,|\,n)\}_{p=0,1,\ldots}$ *converges linearly, then, as* $p \to \infty$, $\log_{10}\{\Lambda^{p+1}/\Lambda^p\}$ *converges to* $\log_{10}\mathrm{r}$.

For the proof see R. S. Pilla's Ph.D. dissertation.

Lemma 2. *If each iteration of an algorithm* B *consists of* q *steps of algorithm* A *that has asymptotic rate* r_{A}, *then the asymptotic rate of algorithm* B *is* $\mathrm{r}_{\mathrm{B}} = \{\mathrm{r}_{\mathrm{A}}\}^q$.

Thus, if we have an algorithm C with $\log \mathrm{r}_{\mathrm{C}}/\log \mathrm{r}_{\mathrm{A}} = q$, then one step of algorithm C provides the same asymptotic increase in the likelihood as do q steps of algorithm A.

5·2. *Convergence criteria*

The following gradient-based stopping criterion has a solid theoretical foundation. It follows from Lindsay (1995, pp. 131–2) that, if we stop when $\sup_j D_\pi(j) \leqslant \varepsilon$, with $\varepsilon = 0{\cdot}005$, where the gradient function $D_\pi(j)$ is defined as

$$D_\pi(j) := \sum_{t=0}^{T} n(t)\left\{\frac{f_j(t)}{g(t\,|\,\pi)} - 1\right\}, \tag{5·1}$$

for all $j = 1, \ldots, m$, then we automatically satisfy the 'ideal stopping criterion':

$$|\log L_{\text{obs}}(\hat{\pi}\,|\,n) - \log L_{\text{obs}}(\pi^p\,|\,n)| \leqslant \varepsilon.$$

The gradient function therefore creates a natural stopping rule for iterative algorithms in

the mixture problem when the final loglikelihood is unknown, although it is more stringent than the ideal stopping rule. In our examples, we instead compute the final loglikelihood to a high degree of accuracy so that we create a rule that is very close to the ideal stopping rule.

In a potentially slow algorithm like EM, a far weaker convergence criterion based on $|\log L_{\text{obs}}(\pi^p|n) - \log L_{\text{obs}}(\pi^{p-1}|n)|$ or on changes in the parameter estimates between iterations can be very misleading, as discussed by Titterington et al. (1985, p. 90); inherently slower algorithms are stopped at smaller likelihood values as the stepwise changes are smaller. See Lindsay (1995, pp. 62–3) for further discussion.

5·3. *Problem of suboptimal solution*

If one uses an algorithm that eliminates components with zero weight parameters at intermediate steps, as discussed in § 3·4, then it is possible that the Newton–Raphson step in paired EM may set a positive weight to zero that belongs to the final solution. If so, removing it permanently results in a suboptimal solution. We show how to use the gradient inequality (5·1) to solve this problem. If the gradient inequality for convergence shows no violator at the specified tolerance, i.e. all j satisfy $D_\pi(j) \leqslant \varepsilon$, then one is done. However, if some support points, π_j, π_{j+1} and π_{j+2}, say, for any $1 < j < m-2$, violate $D_\pi(j) \leqslant \varepsilon$ and currently have zero weights, then we must put them back into the active or positive set at the next iteration. This can be accomplished, while maintaining the monotonicity of the loglikelihood, using either of the following two approaches, compared numerically in § 6·2.

Approach 1. Let $f^*(t) = \frac{1}{3}f_j(t) + \frac{1}{3}f_{j+1}(t) + \frac{1}{3}f_{j+2}(t)$. Then one takes one step towards maximising $\sum n(t) \log\{(1-\eta)g(t|\pi) + \eta f^*(t)\}$, where $0 \leqslant \eta \leqslant 1$, with respect to η using one Newton–Raphson step. We then distribute the η equally among π_j, π_{j+1} and π_{j+2}, and multiply the rest of the π parameters by $(1-\eta)$ to meet the constraint $\sum \pi_j = 1$.

Approach 2. Suppose that π_{j+1} has the largest positive gradient value among the violators. Then insert only the π_{j+1} in the next iteration by finding the mass using one Newton–Raphson step, as in the above procedure. It may well be that this π_{j+1} pushes the π_j and π_{j+2} to zeros in a later iteration, and hence one never needs to put them back into the active set of the algorithm.

5·4. *Counting scheme*

Counting the number of iterations would lead to unfair comparisons because of the varying degrees of numerical complexity as well as the different cycle lengths of the methods. We will therefore count the number of parameter updates, N_{up}, that occur. Thus one EM iteration updates all parameters once, and so is one update. A rotation cycle consists of two updates. In one hierarchical cycle all parameters are effectively updated twice. Hence we will count a single hierarchical cycle as equal to two updates. Finally, a rotated-hierarchical cycle is counted as four parameter updates. We report an effective rate per update by taking the Jth root of the rate per iteration, where J is the number of updates per iteration. We note that the number of updates that each algorithm can undertake in a minute of CPU time is not constant across algorithms. However, it does provide a consistent internal clock independent of programming efficiency. We also provide CPU time for our methods.

6. Examples

6·1. *Sibship data and binomial mixtures*

Note that in this and next examples one would ordinarily fit parameters continuously, not on a grid as done here. However, one could use our grid solution as a starting value for a continuous solution, as done by Böhning et al. (1992). All the computations were done using AEM-FMIX, a special set of Fortran programs for alternative EM in finite mixtures that is available from R. S. Pilla.

The dataset of Table 2 in Lindsay & Roeder (1992) gives the number of male children among the first 12 children in 6115 sibships of size 13, collected in Saxony, Germany. The last born child was ignored in an effort to minimise the possible effects of stopping rules. From the discussion in Lindsay & Roeder (1992), one might model the data as being from a mixture of binomials with $n = 12$ and where θ, the constant probability of having a male child, has an unknown latent distribution.

To put the problem in our setting, we assume that θ_j $(j = 1, \ldots, d+1)$, for $d = 31$, arise from a discrete distribution with 32 support points on $[0, 1]$, namely $\{0, \frac{1}{31}, \ldots, \frac{31}{31} = 1\}$, with corresponding unknown weights $\pi_1, \ldots, \pi_{32}$. In order to investigate the effect of increasing the number and closeness of densities, we also consider the 64-point grid $\{0, \frac{1}{63}, \ldots, \frac{63}{63} = 1\}$, with unknown weights $\pi_1, \ldots, \pi_{64}$.

From uniform initial values, the final loglikelihood values of

$$\log L_{\text{obs}}(\hat{\pi} \mid n) = \log L^{\max} = \begin{cases} -12\,490{\cdot}8214, & \text{for } m = 32, \\ -12\,490{\cdot}7804, & \text{for } m = 64, \end{cases}$$

were found with a high degree of accuracy by running one of the fastest algorithms, hierarchical EM, for a large number of parameter updates. These were used to assess the accuracy of different algorithms at any stage. The maximum likelihood estimates of the weight parameters at the corresponding support points, for the mixture of 32 and 64 densities, are

$$\begin{pmatrix} \hat{\pi} \\ \theta \end{pmatrix} = \begin{pmatrix} 0{\cdot}0001 & 0{\cdot}0051 & 0{\cdot}512 & 0{\cdot}3098 & 0{\cdot}0015 & 0{\cdot}0211 & 0{\cdot}1492 & 0{\cdot}0001 \\ \frac{6}{31} & \frac{7}{31} & \frac{15}{31} & \frac{16}{31} & \frac{18}{31} & \frac{19}{31} & \frac{20}{31} & 1 \end{pmatrix},$$
$$\begin{pmatrix} \hat{\pi} \\ \theta \end{pmatrix} = \begin{pmatrix} 0{\cdot}0068 & 0{\cdot}6969 & 0{\cdot}1091 & 0{\cdot}0320 & 0{\cdot}0956 & 0{\cdot}0596 & 0{\cdot}0001 \\ \frac{14}{63} & \frac{31}{63} & \frac{32}{63} & \frac{39}{63} & \frac{40}{63} & \frac{41}{63} & 1 \end{pmatrix}, \tag{6·1}$$

respectively. The fact that the weights are zero on most of the grid means that the algorithms must push the estimates to the boundary of the parameter space, a least favourable case for EM.

The initial and long-run behaviour of the algorithms. Figure 1(a), a plot of the log residual, $\log_{10} \Lambda^p$ versus $N_{\text{up}} = p$, shows the behaviour of the algorithm over the first 1000 parameter updates. The logarithmic scaling of the vertical axis is used because a linearly convergent algorithm will become linear on this scale as $p \to \infty$; see Lemma 1. The y-axis scale has been converted back to Λ-units. For example, one can see that hierarchical EM has attained a residual of $\Lambda^p = 0{\cdot}1$ at $p \doteq 300$. All the algorithms are extremely fast in the initial stages, and are within about one unit of the final loglikelihood, that is $\Lambda^p = 1$, by $p = 200$. However, the cyclical versions of paired EM hold their initial acceleration longer, and even though they do start to level off at $p \doteq 600$ they are still moving downhill faster over the whole range.

Figure 1(b) shows the behaviour of the algorithms over the whole range of the 209 600

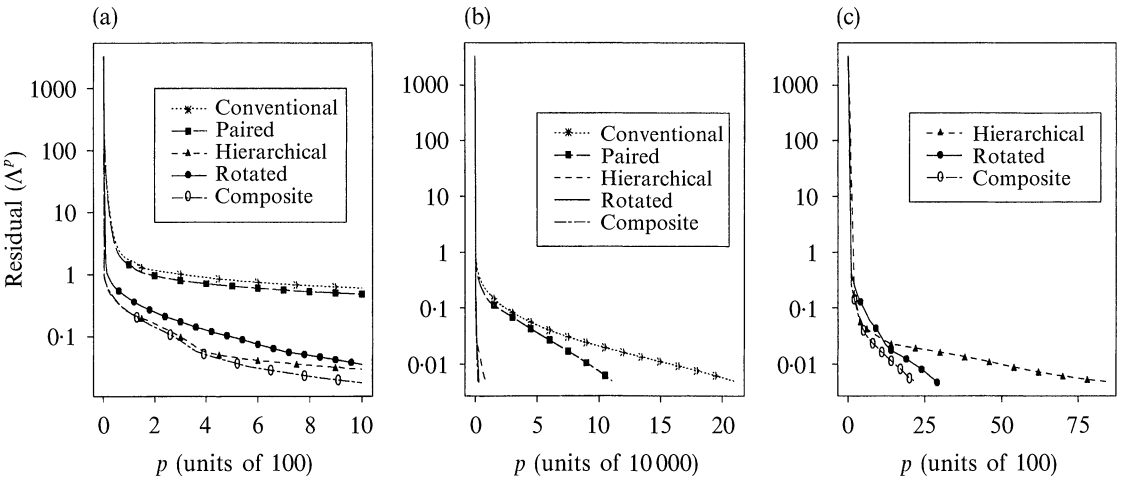

Fig. 1. Residual reached by the algorithms for mixtures of 64 densities plotted against number of updates $N_{up} = p$ for conventional, paired, hierarchical, rotated and composite EM algorithms.

parameter updates needed to ensure that conventional EM had converged to a residual of $\Lambda = 0{\cdot}005$, based on the ideal stopping rule discussed in § 5·2. The plot shows three things: (i) the remarkable early speed together with the nearly linear behaviour in the tails, (ii) the smaller the residual is, the more important the selection of the algorithm becomes, and (iii) the three cyclical versions of basic paired EM converge much faster than conventional or paired EM. From the criss-crossing behaviour of the curves in Fig. 1(c), it is clear that, although hierarchical EM is faster than rotated EM initially, at smaller levels of residual it is slower.

Effect of increasing the number of overlapping densities. Table 1 demonstrates the effects on the algorithms of increasing the number and closeness of densities, from $m = 32$ to $m = 64$. The conventional, paired and hierarchical EM's required nearly four times as many updates at the smallest level of residual, $\Lambda = 0{\cdot}01$, with only a doubling of the number of parameters. However, the rotated and composite EM's showed only about a two-fold increase in updates at the same residual. On the whole, for all the methods, the larger the residual that was required, the smaller was the degradation in performance that occurred in increasing the number of densities.

Table 1. *Parameter updates, N_{up}, needed for various levels of residual, Λ*

	$m = 32$			$m = 64$		
Method	$\Lambda = 1$	$\Lambda = 0{\cdot}1$	$\Lambda = 0{\cdot}01$	$\Lambda = 1$	$\Lambda = 0{\cdot}1$	$\Lambda = 0{\cdot}01$
Conventional EM	255	16900	46900	312	24100	161400
Paired EM	122	7500	23700	183	18100	99600
Hierarchical EM	4	192	1200	4	302	5200
Rotated EM	14	416	1300	17	499	2300
Composite EM	4	160	1000	4	276	1562

Performance based on several criteria. We consider three key measures for comparison of the algorithms, namely the number of parameter updates, N_{up}, required to reach the desired residual of $\Lambda = 0{\cdot}005$, the corresponding CPU time relative to that of EM and the rate of convergence of each algorithm. The results are summarised in Table 2.

Table 2. *Comparisons at the residual level of* $\Lambda = 0{\cdot}005$

Method	$m = 32$			$m = 64$		
	Improvement			Improvement		
	N_{up}	Speed	|Log rate|	N_{up}	Speed	|Log rate|
Conventional EM	1	1	0·0000240	1	1	0·0000057
Paired EM	2	7	0·0000683	2	6	0·0000276
Hierarchical EM	30	42	0·0004450	25	29	0·0001035
Rotated EM	26	105	0·0002726	72	185	0·0004260
Composite EM	49	70	0·0010768	95	92	0·0006340

In Table 2 the N_{up} and speed improvement columns give the relative improvement of each method over EM with respect to N_{up} and the inverse of the CPU time of the algorithm at hand. Time was measured on a SUN SPARC station 4 with 32 MB RAM and 110 MHz clock speed. The conventional EM algorithm took 8·43 minutes with $N_{up} = 59\,000$, for $m = 32$, and 40·62 minutes with $N_{up} = 209\,600$, for $m = 64$, whereas rotated EM took only 4·8 seconds with $N_{up} = 2300$, for $m = 32$, and 13·2 seconds with $N_{up} = 2900$, for $m = 64$. Rotated EM is a clear winner with respect to speed.

In Table 2 the |Log rate| column gives the asymptotic log rate of convergence, which is approximated using an estimate of the slope of the terminal linear portion of the corresponding curve in Fig. 1(b). The slope was calculated using number of updates on the horizontal axis. If one wanted the log rate per true iteration instead, one would multiply this number by J, the number of parameter updates per iteration. We estimate the relative effectiveness of a single update using Lemma 2. For example, for $m = 64$ it takes about 111 EM updates, i.e. approximately 0·0006340/0·0000057, to improve the log-likelihood by as much as a single composite EM update. The clear overall winner in rate was composite EM.

6·2. *Galaxy data and normal mixtures*

Consider the data from Roeder (1990) consisting of velocities of 82 galaxies from 6 well-separated conic sections of space. As suggested by Roeder, natural candidates for examining the distribution of the data are finite normal mixtures.

The set-up for this example is exactly the same as before, with the exception that we use a grid of 64 support points such that $\theta_j \in \{10{\cdot}0, 10{\cdot}38, \ldots, 33{\cdot}56, 33{\cdot}94\}$. This grid is chosen in such a way that the number of densities, m, is large and they are highly overlapping. The endpoints of the grid are chosen to include the estimates of the parameters obtained by Roeder. An equal-spaced grid is chosen for simplicity. The standard deviation, σ, is assumed to be 0·95, as obtained by Roeder using the method of least squares crossvalidation.

Fitting the mixture of 64 normals to the data resulted in the maximum likelihood estimates of the π parameters at the corresponding support points as

$$\begin{pmatrix} \hat{\pi} \\ \theta \end{pmatrix} = \begin{pmatrix} 0{\cdot}085 & 0{\cdot}025 & 0{\cdot}397 & 0{\cdot}060 & 0{\cdot}282 & 0{\cdot}078 & 0{\cdot}036 & 0{\cdot}001 & 0{\cdot}013 & 0{\cdot}024 \\ 10 & 16{\cdot}08 & 19{\cdot}88 & 20{\cdot}26 & 22{\cdot}92 & 23{\cdot}68 & 26{\cdot}34 & 26{\cdot}72 & 32{\cdot}8 & 33{\cdot}18 \end{pmatrix}.$$

Table 3 gives a summary of the performance of the various algorithms. Composite EM is incredibly fast in this case. The conventional EM algorithm took $N_{up} = 20\,400$ and 18·08 minutes, whereas composite EM took only $N_{up} = 56$ and 3·41 seconds to reach

$\log L_{\text{obs}}(\hat{\pi}\,|\,n) = -199{\cdot}03604156$. It is clear that it takes about 968 EM updates, i.e. approximately 0·111374/0·000115, to improve the loglikelihood by as much as a single composite EM update. To see the effect of varying σ^2, we considered the cases of (i) $\sigma^2 = 4(0{\cdot}95)^2$ and (ii) $\sigma^2 = 2(0{\cdot}95)^2$. The improvement of composite EM over conventional EM in parameter updates and speed was respectively 406 and 300 times in case (i), and 75 and 82 times in case (ii). In the latter case, the densities are less overlapping and EM itself is fast, so the improvement is not remarkably high.

Table 3. *Comparisons at the residual level of* $\Lambda = 0{\cdot}005$ *for the normal mixtures*

Method	Improvement N_{up}	Improvement Speed	|Log rate|
Conventional EM	1	1	0·000115
Paired EM	3	9	0·000574
Paired EM (Z)	81† (85‡)	363† (362‡)	0·029474†
Hierarchical EM	291	272	0·089692
Rotated EM	68	173	0·022472
Rotated EM (Z)	255† (237‡)	583† (232‡)	0·071263†
Composite EM	364	318	0·111374

(Z) refers to results for zero-elimination with † for Approach 1 and ‡ for Approach 2, defined in § 5·3.

Table 3 also presents results for zero-elimination for two of the algorithms. To avoid the problem of a suboptimal solution we implemented the gradient check using both the approaches discussed in § 5·3. We checked the gradient inequality at the 10th stage and then at the end of the iterative scheme for Approach 1, and at every 25th and 5th stages for the paired and rotated EM's respectively for Approach 2. The convergence behaviour of the observed loglikelihood sequence in Approach 2 is sensitive to the frequency of the gradient check; the frequencies presented here were empirically optimal among a number of different ones. Hence we recommend Approach 1. From Table 3 it is clear that zero-elimination improves the rate of convergence of paired EM by forty-fold while that of rotated EM has a four-fold increase. Once we have reduced to pairing only nonzero components, the difference between the two algorithms is less than two-fold.

6·3. *A comparison with continuous support EM*

The standard continuous support EM algorithm requires the specification of initial values for the number and location of the support points. If we are seeking the nonparametric mixture estimator, one has to allow for the possibility that too few or too many support points have been used in the initial set. If the initial values, including the number of points used, are close to the right solution, then the algorithm can be very fast and accurate relative to grid-based EM. Unfortunately, the algorithm can never increase the number of support points, so this choice is critical. Also, the optimisation problem solved by the grid-based approach has a unique optimum, which in turn generates an estimated number of components. However, the continuous support EM can converge to a suboptimal mode. In addition, there are mixture problems in which the latter EM algorithm does not apply but the methods of this paper do, as cited in § 1.

We will illustrate the above points through examples. If in the sibship data we started with equal weights at 0·25, 0·50, 0·75 and 1·0, the continuous support EM algorithm reaches

a maximum of $-12\,490{\cdot}7698$ after 10 599 iterations in 19·34 seconds, whereas fixed support mixture EM with 64 points took over 209 600 iterations in 40·62 minutes to reach the same maximum. However, continuous support EM required a 'lucky choice' of initial values and number of components. If instead we started continuous support EM with equal weights on the 64-point grid $\{0, \frac{1}{63}, \ldots, \frac{62}{63}, 1\}$, then it took 6500 updates in 2·52 minutes to reach the same maximum, whereas rotated EM took 2900 updates in 13·2 seconds.

Turning to the normal example, if we start continuous support EM with equal weights at 9·9, 17, 20, 23, 27 and 33, corresponding approximately to the six modes in Fig. 5(c) of Roeder (1990), we quickly reach the maximum of $-198{\cdot}6336$ after 24 iterations in 0·32 seconds. However, if we started with equal weights at 10, 15, 20, 25, 30 and 35, a suboptimal solution of $-206{\cdot}5676$ is reached.

This multimodal example illustrates that, if the number of support points is fixed in advance, then one could have competing explanations, i.e. modes, for the data structure at hand. One may ask whether having multimodal likelihoods provides scientific advantages over the nonparametric maximum likelihood approach which corresponds to a unimodal likelihood with no competing modal explanations. The fitted solution in the nonparametric mixture problem can create competing explanations for the data in a number of ways (Lindsay, 1995, p. 60). For example, one can create more parsimonious models by setting the small π parameters to zero and by merging nearby components, giving a variety of models with similar explanatory power. This would seem to be quicker, and may well be more effective, than searching for alternative modes using multiple start values and differing numbers of support points.

In the light of these characteristics, if one wants a continuous solution with a flexible number of support points, a fast algorithm could be constructed by starting with rotated EM on a grid, and then using the resulting solution as starting values for continuous EM. Such an approach helps to protect against using an incorrect number of points or finding suboptimal solutions.

7. Discussion

The alternative EM methods proposed in this paper can be generalised to a new problem by constructing a class of data augmentations, each smaller than the standard one, such that

(i) the EM algorithm for each augmentation is relatively simple numerically,
(ii) each augmentation has some direction or directions in the parameter space through which it moves quickly relative to the standard algorithm, and
(iii) the set of directions in which acceleration occurs spans the parameter space, so that rotating through the augmentation has a synergistic effect.

It is a bit hard to provide a general methodology for such a construction; it is more a question of insight and creativity.

The examples considered in this paper were designed to have many overlapping densities with most of the weight parameters on the boundary of the parameter space to create an ideal situation for improving the EM; we would not expect substantial improvement over the conventional EM when the densities were well separated and/or most of the weight parameters were positive. In this case, the cyclic versions of the basic paired EM converged many times faster than the EM in the high-dimensional mixtures of binomial and normal. Similar improvements were observed in R. S. Pilla's Ph.D. dissertation for the mixtures of 32 Poissons.

The performance of rotated EM with or without zero-elimination was outstanding. Our general recommendations are as follows. We note that improvement by a factor of 500 may not be significant for a single computation, but for application of the bootstrap or a simulation experiment it could mean the difference between one hour and 20 days. One might prefer rotated EM for its simplicity, but, if most of the weight parameters are zeros, combining it with zero-elimination can make it a clear winner in terms of the rate and speed. The robust efficiency of hierarchical EM might make it worth implementing for large-scale repeated usage.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the constructive comments of the editor, an associate editor and two referees. This research was supported by the National Science Foundation Grants. Part of the work was done while R. S. Pilla was at the National Institutes of Health, Bethesda, MD, U.S.A. and was funded by a National Institutes of Health Fellowship.

REFERENCES

BÖHNING, D., DIETZ, E. & SCHLATTMANN, P. (1998). Recent developments in computer-assisted analysis of mixtures. *Biometrics* **54**, 525–36.

BÖHNING, D. & LINDSAY, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Ann. Inst. Statist. Math.* **40**, 641–63.

BÖHNING, D., SCHLATTMANN, P. & DIETZ, E. (1996). Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika* **83**, 462–6.

BÖHNING, D., SCHLATTMANN, P. & LINDSAY, B. G. (1992). Computer-assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics* **48**, 283–303.

CELEUX, G., CHAUVEAU, D. & DIEBOLT, J. (1996). Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Statist. Comp. Simul.* **55**, 287–314.

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.

JAMSHIDIAN, M. & JENNRICH, R. I. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *J. R. Statist. Soc.* B **59**, 569–87.

LANGE, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. R. Statist. Soc.* B **57**, 425–37.

LINDSAY, B. G. (1983). The geometry of mixture likelihoods: A general theory. *Ann. Statist.* **11**, 86–94.

LINDSAY, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*, NSF-CBMS Regional Conference Series in Probability and Statistics, **5**. Hayward, CA: Institute of Mathematical Statistics.

LINDSAY, B. G. & ROEDER, K. (1992). Residual diagnostics for mixture models. *J. Am. Statist. Assoc.* **87**, 785–94.

MCLACHLAN, G. J. & KRISHNAN, T. (1997). *The EM Algorithm and Extensions.* New York: Wiley.

MENG, X. L. & VAN DYK, D. A. (1997). The EM algorithm – An old folk sung to a fast new tune (with Discussion). *J. R. Statist. Soc.* B **59**, 511–67.

NEAL, R. M. & HINTON, G. E. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, Ed. M. I. Jordan, pp. 355–68. Cambridge, MA: MIT Press.

PILLA, R. S. & LINDSAY, B. G. (1996). Faster EM methods in high-dimensional finite mixtures. In *Proc. Statist. Comp. Sect.* pp. 166–71. Alexandria, VA: American Statistical Association.

ROEDER, K. (1990). Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *J. Am. Statist. Assoc.* **85**, 617–24.

ROEDER, K., DEVLIN, B. & LINDSAY, B. G. (1989). Application of maximum likelihood methods to population genetic data for the estimation of individual fertilities. *Biometrics* **45**, 363–79.

SILVEY, S. D. (1980). *Optimal Design – An Introduction to the Theory for Parameter Estimation.* New York: Chapman and Hall.

TITTERINGTON, D. M., SMITH, A. F. M. & MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* New York: John Wiley.

VARDI, Y., SHEPP, L. A. & KAUFMAN, L. (1985). A statistical model for positron emission tomography (with Discussion). *J. Am. Statist. Assoc.* **80**, 8–37.

[*Received October* 1998. *Revised September* 2000]