cyclic methods, rotation and hierarchical, greatly accelerated the algorithm, but the much simpler rotation cycle was also more effective. Indeed there was no significant improvement in rate when hierarchical cycles were added to the rotation cycles. With the DEM based methods, we see the same basic results as in REM based methods but now we find that the improvements over CEM are not nearly as large as for REM. From this we conclude that substantial gains over DEM methods are possible from the small amount of computational effort needed to implement the Newton-Raphson algorithm in the REM based methods. FORTRAN programs may be obtained by contacting the authors.

The preliminary numerical testing on simulated data suggest that the general methods are superior to the CEM. Also there is a dramatic improvement when the rotation cycles and rotated hierarchical cycles were added to the REM method. The methods seem to show striking gains with out any additional computational increase especially when the component densities are poorly separated and/or when some of the weight parameters lie on the boundary of the parametric space. It appears that the improvements of the general methods are remarkably significant in high dimensional mixtures. Among issues we plan to investigate further are application to large scale problems, extensions to unknown components, and greater understanding of numerical properties.

Table 1: Comparison of Several Methods for Different m.

Method	m = 32			m = 64		
	Rate	Improvement		Rate	Improvement	
		Itn	Time		Itn	Time
DEM	0.990	2.27	7.88	0.996	1.42	3.12
REM	0.994	1.57	7.08	0.996	1.90	5.87
$\mathbf{DEM^h}$	0.908	9.67	6.11	0.998	3.25	1.42
$\mathbf{REM^h}$	0.848	59.0	42.15	0.945	49.90	29.01
$\mathbf{DEM^r}$	0.605	4.65	12.58	0.605	8.19	13.41
$\mathbf{REM^r}$	0.605	25.65	105.38	0.605	72.28	184.64
$\mathbf{DEM^{rh}}$	0.605	6.41	2.53	0.605	12.33	2.89
$\mathbf{REM^{rh}}$	0.605	98.33	70.25	0.605	190.55	92.32

References

Böhning, D. and Lindsay, B. G. (1988). Monotonicity of quadratic-approximation algorithms. *Ann. of the Inst. of Stat. Math.*, 40:641–663.

Böhning, D., Schlattmann, P., and Dietz, E. (1996). Interval censored data: A note on the nonparametric maximum likelihood estimator of the distribution function. *Biometrika*, 83:462–466.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. of Roy. Statist. Soc. Ser. B*, 39:1–22.

Lange, K. (1995). A gradient algorithm locally equivalent to the EM algorithm. *J. Roy. Statist. Soc. Ser. B.*, 57:425–437.

Lindsay, B. G. (1984). Optimal EM algorithms in polynomial problems. Technical Report 47, The Pennsylvania State University, University Park, PA-16802.

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

Lindsay, B. G. and Roeder, K. (1992). Residual diagnostics for mixture models. *J. Amer. Stat. Assoc.*, 87:785–794.

Pilla, R. S. and Lindsay, B. G. (1996). Alternative EM algorithms in high-dimensional finite mixtures. Technical Report 96-2, The Pennsylvania State University, University Park, PA-16802.

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

Vardi, Y., Shepp, L. A., and Kaufman, L. (1985). A statistical model for positron emission tomography. *J. Amer. Statist. Ass.*, 80:8–37.

In the second EM step of the cycle, we would use the block EM on the blocks $A_1 = (f_1, f_2), A_2 = (f_3, f_4), A_3 = (f_5, f_6)$, and $A_4 = (f_7, f_8)$ to accelerate the relative weights $\beta_1 = (\pi_1 + \pi_2)/(\pi_1 + \cdots \pi_4)$ and $\beta_2 = (\pi_5+\pi_6)/(\pi_5+\cdots\pi_8)$. Finally in the third EM-step we use the blocks $A_1 = (f_1, f_2, f_3, f_4), A_2 = (f_5, f_6, f_7, f_8)$ to accelerate the parameter $\gamma = (\pi_1 + \cdots + \pi_4)$. Since $(\alpha_1, \alpha_3, \alpha_5, \alpha_7, \beta_1, \beta_2, \gamma)$ is a one-to-one reparametrization of the π's, this three step cycle updates and accelerates throughout the parameter space. (There is no concern about convergence here, as the first EM step updates all parameters.) Such a hierarchical scheme is easily extended to any number of densities, one simply takes any pairing scheme and at each successive stage of the cycle, each pair of blocks in the previous EM-step becomes merged into a single block at the next level. This scheme continues hierarchically until there are two blocks left. The scheme is easiest to implement when there are 2^k densities, in which case there are k EM-steps within each cycle.

When we implement our data augmentations with hierarchical EM cycles, we will denote them $\mathbf{DEM^h}$ and $\mathbf{REM^h}$ respectively. Moreover, we can obviously create an even larger cyclic scheme by rotating the densities between each application of the hierarchical cycle. These methods will be denoted by the superscript **rh** for "*rotated hierarchical cycles*"

5. Results and Conclusions

We have now introduced two new data augmentation schemes for the fixed component model as well as two different ways to construct cyclic EM structures from each of them. In this section we do a numerical case study to see the relative effectiveness of these various methods.

For our comparison, we will fit a mixture model to the data set of Table 2 given in Lindsay & Roeder (1992) that gives the number of male children in 6115 sibships of size 12, collected in Saxony, Germany. As noted by them, one could model the children of a seven couple as being born with independent sex determination, like a sequence of Bernoulli trials with constant probability p of having a male child. If so, the number of male children X in a family with 12 children would be distributed as a $Bin(12, p)$ random variable. If probability of a male birth p varies from couple to couple, then we could associate with each couple a latent parameter p_j representing their propensity to have male children. The resulting data would be from a mixture of binomials with $n = 12$ and p from an unknown latent distribution.

Setting: To put the problem in our setting, we assume the p_j arise from a discrete distribution with 32 support points on $[0, 1]$ namely $\{0, 1/31, 2/31, \ldots, 31/31 = 1\}$, 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 $\{0, 1/63, 2/63, \ldots, 63/63 = 1\}$, with unknown weights $\pi_1, \ldots, \pi_{64}$.

In order to compare the algorithms, we needed a stopping rule. This was set by taking very long runs with the fastest algorithm and finding the $l(\widehat{\boldsymbol{\pi}})$, the final log likelihood value. Each of the other algorithms was then started with uniform weights and run until the $l(\widehat{\boldsymbol{\pi}}) - l(\boldsymbol{\pi}_{\text{stop}}) \leq 0.005$, when it was said to "converge". As Lindsay (1995) suggested, this likelihood deviation represents what is needed for reasonable inferential accuracy. The final solution we arrived at from our best method, REM^h, has many weights at zero, corresponding to a least favorable case for CEM.

For the $m = 32$ and $m = 64$ densities case, the CEM algorithm took 59,000 and 209,200 iterations with CPU times of 8.43 minutes and 40.62 minutes respectively to converge to the solution at which the observed log likelihoods, l_{obs}, are -12490.8264 and -12490.78547 respectively.

It is somewhat difficult to compare the various algorithms developed here in terms of the "number of iterations" as there are different degrees of numerical complexity as well as different cycle lengths. We will focus here on the CPU time relative to that of the CEM as a yardstick. For the reader's information, we will also present the relative number of iterations, where for simplicity we call each complicated hierarchical cycle a single "iteration". Since the rotation cycles are simply two simple EM's back-to-back, it is most natural for comparisons to consider each cycle as two iterations. As a second yardstick we also present the "rate of convergence" of the log likelihood, which is the empirical rate of convergence of the log likelihood function for each method calculated from every 100th iteration that measures the gain at each step relative to the gain in the step before.

In Table 1, we examine the various methods. The "*Improvement*" column gives the improvement of each method over the CEM w. r. t. the iterations and time respectively. We note that the basic REM is less than twice as fast as the CEM in number of iterations, but is about five times faster in CPU time. This occurs because the implemented REM can automatically set component weights equal to zero in the Newton-Raphson step, thereby removing them from further computation. We note also that both

simple. The lack of an explicit solution to the maximization is compensated for by the quadratic convergence of NR. A practical question arises as to how many NR steps to take within each EM-step. Lange [1995, Proposition 1] has shown that the EM's rate of convergence is independent of the number of NR steps within each EM-step, so one-step is sufficient. We adopt this method here.

4. Two New Complete Data Cycles

Both of our new data augmentation schemes depend upon a selected pairing of densities $[(f_1, f_2), (f_3, f_4), \ldots]$. The corresponding EM algorithm then accelerates the determination of the relative weights "α" within each pair, while providing no improvements in the other $\pi^\star$ parameters. We next develop two "*cyclic missing data*" algorithms that will enable us to extend the acceleration directions to the whole parameter space.

4.1 Rotation Cycles

Our simplest idea is to simply choose a cycle of pairings of densities such that we "*rotate*" through a sequence of possible pairings of similar densities whose relative weightings would benefit from acceleration. Within each pairing scheme, we simply apply one step of the DEM or REM, then move to the next pairing scheme. Suppose there are four densities, and adjacent pairs were most similar. In this case we would pair: (f_1, f_2) and (f_3, f_4) in odd iteration and (f_2, f_3) and (f_4, f_1) in even iteration or vice versa. As we will see, this very simple two step cycle is highly effective. Intuitively, it is because the transferral of mass between 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 improvement of mass to any one density that will receive large final weight.

When we implement our two basic data augmentation schemes with two-step rotation cycles, they will be denoted $\mathbf{DEM^r}$ and $\mathbf{REM^r}$ for *rotation cycles*.

4.2 Hierarchical Cycles

Our second cyclic algorithm is based upon the observation that we can use either of our data augmentation schemes to accelerate the determination of the relative weights of any set of pairings of blocks of densities. We illustrate this *partial block EM* principle first, before turning it into a cyclic approach.

1. The Block EM: Let the *blocks of indices* be defined by $A_1 = \{1, \ldots, a_1\}, A_2 = \{a_1 + 1, \ldots, a_2\}, \cdots, A_{B-1} = \{a_{B-2} + 1, \ldots, a_{B-1}\}$, and $A_B = \{a_{B-1} + 1, \ldots, a_B \equiv m\}$. The blocks will be paired together, $(A_1, A_2), (A_3, A_4), \ldots$, and so forth. Once again, it is useful to reparametrize the problem: Let $\pi_i^\dagger = \sum_{j \in A_i} \pi_j$ for all $i = 1, \ldots, B$,

$$\pi_\rho^\star = \sum_{j \in A_{2\rho-1}} \pi_j + \sum_{j \in A_{2\rho}} \pi_j = \pi_{2\rho-1}^\dagger + \pi_{2\rho}^\dagger,$$

$$\alpha_{2\rho-1} = \frac{\pi_{2\rho-1}^\dagger}{\pi_{2\rho-1}^\dagger + \pi_{2\rho}^\dagger} = \frac{\pi_{2\rho-1}^\dagger}{\pi_\rho^\star},$$

and $\alpha_{2\rho} = 1 - \alpha_{2\rho-1}$ for all $\rho = 1, \ldots, B/2$. Thus $\pi^\dagger$ is the sum of all the π's in a given block and $\pi^\star$ is the sum of the two $\pi^\dagger$'s in adjacent blocks.

Construct the *relative block densities* using $P_i(t) = \sum_{j \in A_i} \alpha_j^\dagger f_j(t)$, where

$$\alpha_j^\dagger = \frac{\pi_j}{\sum_{j \in A_i} \pi_j} = \frac{\pi_j}{\pi_i^\dagger}. \tag{6}$$

When we apply the EM in this setting, the parameters $\alpha_j^\dagger$ will be treated as fixed, and so the $P_i(t)$ will play the role of $f_j(t)$ in the algorithm. Notice that the original weight parameters π_j for all j can be retrieved using $\pi_j = \alpha_j^\dagger \cdot \pi_i^\dagger = \alpha_j^\dagger \cdot \alpha_i \cdot \pi_\rho^\star = \frac{\pi_j}{\pi_i^\dagger} \cdot \frac{\pi_i^\dagger}{\pi_\rho^\star} \cdot \pi_\rho^\star$.

After reparametrization, the mixture density can be rewritten as

$$g\left(t \middle| \boldsymbol{\pi}^\star, \boldsymbol{\alpha}\right) = \sum_{i=1}^{B} \pi_\rho^\star \cdot \left\{\alpha_{2\rho-1} P_{2\rho-1}(t) + \alpha_{2\rho} P_{2\rho}(t)\right\},$$

where $\boldsymbol{\pi}^\star = \left(\pi_1^\star, \ldots, \pi_{B/2}^\star\right)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_B)$. The Block-EM requires us to: (1) Apply the REM or DEM to the family of paired block density representation: $\sum_i \pi_\rho^\star \cdot \{\alpha_{2\rho-1} P_{2\rho-1}(t) + \alpha_{2\rho} P_{2\rho}(t)\}$ to update the parameters $\pi_\rho^\star$'s and α_i's, conditional on the $\alpha_j^\dagger$'s that are "*hidden*" in $P(t)$'s. (For REM, we use Newton-Raphson on the α's.) (2) After obtaining the α_i^{new} for all $i = 1, \ldots, B$ and $\pi_\rho^{\star\,\text{new}}$ for all $\rho = 1, \ldots, B/2$ through the EM-step, update the original π_j's using $\pi_j^{\text{new}} = \alpha_j^{\dagger\,\text{old}} \cdot \alpha_i^{\text{new}} \cdot \pi_\rho^{\star\,\text{new}}$

$$= \frac{\pi_j^{\text{old}}}{\pi_i^{\dagger\,\text{old}}} \cdot \alpha_i^{\text{new}} \cdot \pi_\rho^{\star\,\text{new}} \quad \text{for all} \quad j.$$

2. The Hierarchical Scheme: Given any mixture density, one can construct a reparametrization in terms of the relative weights of pairs of blocks of densities. We illustrate how one could do this with eight densities. The first set of pairs would be $(f_1, f_2), (f_3, f_4), (f_5, f_6)$, and (f_7, f_8), and we would use DEM or REM to obtain acceleration in the relative weights $\alpha_1 = \pi_1/(\pi_1 + \pi_2), \alpha_3 = \pi_3/(\pi_3 + \pi_4), \alpha_5 = \pi_5/(\pi_5 + \pi_6)$, and $\alpha_7 = \pi_7/(\pi_7 + \pi_8)$.

3.2 Multicomponent Delta EM

A simple extension of the DEM to the multicomponent case is achieved by pairing component densities and using an artful reparametrization. For example, with five densities we may write the mixture density as $\pi_1^\star(\alpha_1 f_1 + \alpha_2 f_2) + \pi_3^\star(\alpha_3 f_3 + \alpha_4 f_4) + \pi_5 f_5$, where $\pi_1^\star = \pi_1 + \pi_2, \alpha_1 = \pi_1/(\pi_1 + \pi_2)$, and so forth. We can then use the delta representation method on the parenthetical terms; for example, we can write $\alpha_1 f_1 + \alpha_2 f_2 = h_1 + \alpha_1 \Delta_1^+ + \alpha_2 \Delta_1^-$ and $\alpha_3 f_3 + \alpha_4 f_4 = h_3 + \alpha_3 \Delta_3^+ + \alpha_4 \Delta_3^-$

This representation leads in turn to complete data contingency table with seven rows, with respective multinomial probabilities $\pi_1^\star h_1, \pi_1^\star \alpha_1 \Delta_1^+, \pi_1^\star \alpha_2 \Delta_1^-, \pi_3^\star h_2, \pi_3^\star \alpha_3 \Delta_3^+, \pi_3^\star \alpha_4 \Delta_3^-$, and $\pi_5 f_5$. The corresponding EM algorithm gives explicit solutions for the new parameter set $\pi_1^\star, \pi_2^\star, \alpha_1$, and α_3. This representation idea can be applied to any number of densities by choosing pairs $(f_1, f_2), (f_3, f_4), (f_5, f_6)$, etc. Once done, it can be turned into a complete data structure, just as was done above in the five component case, and from it one can generate a DEM-step.

In an intuitive sense, it is clear that we will gain the greatest improvement over the CEM in the updates of the α-terms, which are the relative weights $\pi_i/(\pi_i + \pi_{i+1})$ assigned to each member of the pair of densities (f_i, f_{i+1}). On the other hand, the $\pi^\star$-weights, $(\pi_1 + \pi_2, \pi_3 + \pi_4, \ldots,)$ are being updated just as in the CEM and so are not accelerated. Thus a crucial as part of this algorithm is that the pairs be chosen to be as similar as possible.

3.3 Partially Missing Data Structure

We now develop a *third data augmentation* mechanism for the mixture problem and the corresponding EM algorithm. It has an advantage over the DEM in that it gives a superior level of theoretical performance. On the other hand, it is inferior to the DEM in that the corresponding EM-likelihood does not have explicit solutions, and so one must use a search algorithm. The missing data structure we use will again depend on the choice of the pairs of densities, say $[(f_1, f_2), (f_3, f_4), \ldots,]$ where the last density f_m is treated separately if m is odd.

Recalling that the complete data for the CEM is the contingency table $\{n_j(t)\}$, we consider the consequence of having the *reduced complete data* $N^\star(t) = [n_1(t) + n_2(t), n_3(t) + n_4(t), \ldots,]$, where the last entry is $n_m(t)$ if m is odd. In this case it is clear we have reduced the missing data by about 1/2, and so expect considerably better performance. However, there is a price to pay that is obvious when $m = 2$: In this case the reduced complete data is just the observed data $n(t)$, and so there is not an explicit solution to the M step, which requires directly maximizing the likelihood (3). On the other hand the Newton-Raphson algorithm for the parameter π is not only quadratically convergent, it is nearly monotonic (Böhning and Lindsay, 1988) and so is highly reliable here.

We will show that even in higher dimensional problems, the reduced complete data, $N^\star$, generates a "nice" optimization problem. We will call $N^\star$ the *Reduced Complete Data* (**RCD**) and label the corresponding EM the **REM**.

3.3.1 The REM Method for Four Component Mixtures

We will illustrate the REM method in the case of four densities; the generalization to more components is obvious. In this case, the CEM algorithm is based on complete data $[n_1(t), \ldots, n_4(t)]$ with $n_1(t) + \cdots + n_4(t) = n(t)$ for $t = 0, \ldots, T$. The RCD is $N^\star(t) = [N_1(t) = n_1(t) + n_2(t), N_2(t) = n_3(t) + n_4(t)]$. In this situation, the complete data likelihood function can be written as $L_{\text{RCD}}(\boldsymbol{\pi}|N^\star)$

$$\propto \prod_{t=0}^{T} \Big[\pi_1 f_1 + \pi_2 f_2\Big]^{N_1(t)} \cdot \Big[\pi_3 f_3 + \pi_4 f_4\Big]^{N_2(t)}.$$

Reparametrization using $\alpha_j = \pi_j/(\pi_j + \pi_{j+1})$ and $\alpha_{j+1} = 1 - \alpha_j$ for odd j gives $L_{\text{RCD}}(\boldsymbol{\pi}^\star, \boldsymbol{\alpha}|N^\star)$

$$\propto \prod_t \underbrace{(\pi_1 + \pi_2)}_{\pi_1^\star}{}^{N_1(t)} \cdot \Big[\alpha_1 f_1 + \alpha_2 f_2\Big]^{N_1(t)} \cdot \underbrace{(\pi_3 + \pi_4)}_{\pi_2^\star}{}^{N_2(t)} \cdot \Big[\alpha_3 f_3 + \alpha_4 f_4\Big]^{N_2(t)}, \quad (5)$$

where $\boldsymbol{\pi}^\star = (\pi_1^\star, \pi_2^\star)$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_4)$ are the new parameters to estimate. Notice that $\pi_1^\star + \pi_2^\star = 1, \alpha_1 + \alpha_2 = 1$, and $\alpha_3 + \alpha_4 = 1$, so we still have three unknowns. Set $\boldsymbol{\alpha}_1 = (\alpha_1, \alpha_2)$ and $\boldsymbol{\alpha}_3 = (\alpha_3, \alpha_4)$.

It is important to note from (5) that the complete data log likelihood, l_{RCD}, decomposes into three pieces, say $A(\boldsymbol{\pi}^\star) + B(\boldsymbol{\alpha}_1) + C(\boldsymbol{\alpha}_3)$ where each piece depends on a single parameter. Hence maximizing $E[l_{\text{RCD}}|\mathbf{n}]$ is equivalent to maximizing each of the $A(\cdot), B(\cdot)$, and $C(\cdot)$ functions w.r.t. $\pi_1^\star, \alpha_1$, and α_3 respectively. See Pilla & Lindsay (1996) for details on the EM algorithm for this missing data.

One of the main attractions of this missing data approach is the way that the parameters separate so that l_{RCD} decomposes into several one-dimensional pieces. Although the solutions are not explicit the Newton-Raphson (NR) computation is relatively

the formulas merely involves setting the range of t to be the observed y_i's, and letting $n(t) = 1$.

2.2 The Conventional EM

The EM algorithm is constructed by taking the log likelihood for the hypothetical complete data $\mathbf{x}$ say, $l_{\text{com}}(\boldsymbol{\pi}|\mathbf{x})$, an initial value $\boldsymbol{\pi}^{(0)}$, and maximizing $E\left[l_{\text{com}}(\boldsymbol{\pi}|\mathbf{x})|\text{observed data}, \boldsymbol{\pi}^{(0)}\right] = Q\left(\boldsymbol{\pi}|\boldsymbol{\pi}^{(0)}\right)$. In our case, the observed data is always the column totals $\mathbf{n} = (n(1), \ldots, n(T))$ but the complete data $\mathbf{x}$ will change with our formulations of the problem. In this missing data representation, the likelihood is given by

$$L_{\text{com}}(\boldsymbol{\pi}|\mathbf{x}) \quad = \quad \prod_{j=1}^{m}\prod_{t=0}^{T}\left[\pi_j f_j(t)\right]^{n_j(t)}. \qquad (2)$$

The **E-step** of the EM algorithm is the computation of the conditional expectation of the complete data log likelihood given the observed data: $E\left[l_{\text{com}}(\boldsymbol{\pi}|\mathbf{x})\Big|\mathbf{n}, \boldsymbol{\pi}^{(p)}\right]$. To compute this, we simply take the complete data log likelihood and replace the $n_j(t)$ with their conditional expectations. We can think of the E-step as filling in the unobserved contingency table with estimates of missing values. The M-step updates the current estimates of $\boldsymbol{\pi}$ by using the relative frequencies in the row marginals. These are the values of $\boldsymbol{\pi}$ maximizing $Q\left(\boldsymbol{\pi}|\boldsymbol{\pi}^{(p)}\right)$. This will be called the *Conventional EM* (**CEM**) *algorithm*. For details see Pilla & Lindsay (1996).

3. Alternative Complete Data Structures

In order to accelerate the EM algorithm while staying within its boundaries, we must find alternative complete data structures that have "*less missing data*" than the CEM. In doing so, we focus on the fact that the CEM is at its worst in determining the relative weights $\frac{\pi_j}{\pi_j+\pi_k}$ and $\frac{\pi_k}{\pi_j+\pi_k}$ given to two densities $f_j(t)$ and $f_k(t)$ that are very similar (Lindsay, 1984).

3.1 The Delta EM Method for Two Component Mixtures

We start by developing a *second data augmentation* that results in what we will call the *Delta EM* (**DEM**). We focus first on the simplified version of (1) where $m = 2$; in this case we will use $\boldsymbol{\pi} = (\pi, \overline{\pi})$ for (π_1, π_2) where $\overline{\pi} = 1 - \pi$. One can write (1) as

$$L(\pi, \overline{\pi}|\mathbf{y}) = \prod_{t=0}^{T}\left[\pi f_1(t) + \overline{\pi} f_2(t)\right]^{n(t)}. \qquad (3)$$

Our new missing data structure is motivated by algebraic considerations. One can write the mixture density $g(\mathbf{y}|\boldsymbol{\pi}) = \pi f_1(\mathbf{y}) + \overline{\pi} f_2(\mathbf{y})$ as (after dropping the arguments for the f, h, and Δ)

$$= \begin{cases} f_1 + \overline{\pi}\,[f_2 - f_1], & \text{if } \; f_1 \le f_2, \\ f_2 + \pi\,[f_1 - f_2], & \text{if } \; f_2 \le f_1, \end{cases}$$

$= h + \pi\Delta^+ + \overline{\pi}\Delta^-$ where $h = \min\{f_1, f_2\}$ and $\Delta = f_1 - f_2$. The functions Δ^+ and Δ^- are the *positive* and *negative parts* of Δ, one of which is zero. We have $|\Delta| = \Delta^+ + \Delta^-$ and $\Delta = \Delta^+ - \Delta^-$, a standard representation of Δ as a difference of two nonnegative functions.

We can turn this representation into a missing data problem as follows. Note that if we have discrete missing data $J_i^\star$ taking on values $1, 2\ldots, r$ then the marginal likelihood of $\mathbf{y}$ is

$$\prod_{t=0}^{T}\left[\sum_{j=1}^{r} P(Y_i = t; J_i^\star = j)\right]^{n(t)}. \qquad (4)$$

If the joint distribution of $(\mathbf{y}, J^\star)$ is specified by $P[Y_i = t, J_i^\star = 1] = h(t), P[Y_i = t, J_i^\star = 2] = \pi \cdot \Delta^+(t)$, and $P[Y_i = t, J_i^\star = 3] = \overline{\pi} \cdot \Delta^-(t)$, then the marginal likelihood for $\mathbf{y}$ is $L(\boldsymbol{\pi}) = \prod_t \left[h + \pi\cdot\Delta^+ + \overline{\pi}\cdot\Delta^-\right]^{n(t)}$. This is just proportional to the original observed data likelihood, so $(\mathbf{y}, \mathbf{J}^\star)$ is a legitimate data augmentation for the mixture problem.

In this case, we can think of the complete data as being a multinomial contingency table with three rows having cell probabilities $h(t)$, $\pi\Delta^+(t)$, and $\overline{\pi}\Delta^-(t)$ respectively. Although it appears that we have somehow generated more missing data than the conventional approach, the row corresponding to h does not depend on the parameters. In effect, the missing data put into this row will be where we gain over the conventional approach, and it will be largest where the two densities are similar (Lindsay, 1984).

With this augmentation the likelihood for the complete data, $L_{\text{Delta}}(\boldsymbol{\pi}|\mathbf{x})$, can be written as follows

$$\propto \prod_{t=0}^{T}\left[h\right]^{n_1(t)} \cdot \prod_{t=0}^{T}\left[\pi\cdot\Delta^+\right]^{n_2(t)} \cdot \prod_{t=0}^{T}\left[\overline{\pi}\cdot\Delta^-\right]^{n_3(t)}.$$

The unobserved counts $[n_1(t), n_2(t),\, n_3(t)]$ have multinomial distribution with the conditional cell probabilities given by $P(J^\star = 1|\mathbf{y} = t) = h/g(t|\boldsymbol{\pi}), P(J^\star = 2|\mathbf{y} = t) = \pi\,\Delta^+/g(t|\boldsymbol{\pi})$, and $P(J^\star = 3|\mathbf{y} = t) = \overline{\pi}\,\Delta^-/g(t|\boldsymbol{\pi})$. The DEM consists of filling-in the missing counts in the E-step and finding the new π in the M-step. See Pilla & Lindsay (1996) for details.

Pilla, R. S. & Lindsay, B. G. (1996), Proceedings of the Statistical Computing Section, pp. 166–171, Alexandria, Virginia. American Statistical Association.

FASTER EM METHODS IN HIGH-DIMENSIONAL FINITE MIXTURES

Ramani S. Pilla and **Bruce G. Lindsay**

Pilla, 326 Thomas Building, Pennsylvania State University, University Park, PA-16802

Key Words: EM algorithm, Finite Mixture Distribution, Maximum Likelihood, Missing Data.

1 Introduction and Motivation

There exists a class of statistical problems which require maximization of a high-dimensional polynomial over the simplex. The paper will be concerned with a general class of maximum likelihood (ML) problems in which it is desired to maximize a mixture likelihood with finitely many known component densities over the set of unknown weight parameters. The maximization problem arises in several statistical contexts, for example: (1) Mixture ML estimation with known component densities (Lindsay, 1995), (2) optimal design (Silvey, 1980), (3) interval censored data (Böhning et al., 1996), and (4) indirectly observed Poison processes including the Positron Emission Tomography (PET) problem (Vardi et al., 1985).

In the mixture case, the convergence of Expectation-Maximization (EM) (Dempster et al., 1977) is extremely slow when the components are similar in their densities and when the ML solution requires some of the weight parameters to be zero, as the algorithm can never reach such a boundary point. We develop and investigate EM-based methods to address this problem.

We note that, a particular feature of our strategy is that we wish to stay within the formal EM-structure, due to its simplicity and reliability when applied in very high-dimensional problems. Our strategy relies on two basic principles:

1. Reducing the Complete Data: A key point regarding the EM algorithm for mixtures is that, in fact, there are a multiplicity of ways to specify the "*missing data*" and so there is not unique EM algorithm, the form of the algorithm depends on the specification of the *augmented data* set, which consists of observed data and hypothetical missing data. (We will call the augmented data set the *complete data.*) In fact we will show that there exist EM algorithms that are strictly superior to the conventional EM algorithm. We consider two such missing data structures here.

2. Cycling between Data Augmentations: As an EM algorithm progresses through the parameter space, it has least favorable directions in which the convergence is inherently slow. These directions depend, in turn, on the augmented data structure. Our second main idea is to construct a sequence of complementary missing data structures such that by cycling through them, we achieve improved acceleration in all directions. We will compare two cyclic methods here.

2 The Mixture Problem

Consider the following optimization problem: given parameters $\boldsymbol{\pi}$ in the unit simplex (i.e., which satisfy the constraints $\sum_j \pi_j = 1, \pi_j \geq 0$), maximize the nth degree polynomial in $\boldsymbol{\pi}$

$$L(\boldsymbol{\pi}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{m} \pi_j f_j(y_i) \right], \qquad (1)$$

where $f_j(y_i) \geq 0$ for $i = 1, \ldots, n$; $j = 1, \ldots, m$, are known constants. In particular, equation (1) is the likelihood function for a random sample $Y_1, \ldots, Y_n$ from a mixture of known component densities $f_j(\mathbf{y})$.

2.1 The Multinomial Likelihood

If the component densities are discrete the sampling is i.i.d., we have a *multinomial likelihood.* We will set up the likelihood equations in this setting, the advantage being that we can interpret the action of the algorithm in terms of filling-in-by-expectation of the unobserved cells of a contingency table.

Suppose we have an i.i.d. sample from a discrete mixture density function $g(t|\boldsymbol{\pi}) = P(Y = t|\boldsymbol{\pi})$ with the sample space $\{0, \ldots, T\}$. Then the observed likelihood of a sample $Y_1, \ldots, Y_n$ can be written as

$$L_{\text{obs}}(\boldsymbol{\pi}|\mathbf{y}) = \prod_{i=1}^{n} g(y_i|\boldsymbol{\pi}) = \prod_{t=0}^{T} \Big[g(t|\boldsymbol{\pi}) \Big]^{n(t)}.$$

In this multinomial case, we can reduce the observed data by sufficiency to the counts $n(t) =$ Number of $\{Y_i = t\}$, and it is clear from the representation of the likelihood that the values $n(0), \ldots, n(T)$ are sufficient statistics for the parameter $\boldsymbol{\pi}$. The missing data will be the variables J_i that identify the component labels. The observed data $n(t)$ are the column marginal totals of an unobserved table of counts, $\{n_j(t)\}$, where the *rows* correspond to the different components j and the *columns* correspond to the possible values t of the variable $\mathbf{Y}$ (see Lindsay, 1995 for details).

Although we derive our formulas in this context for their intuitive appeal, all our results are applicable to continuous density problems. The change in