Statistica Sinica **10**(2000), 203-225

NESTING EM ALGORITHMS FOR COMPUTATIONAL EFFICIENCY

David A. van Dyk

Harvard University

Abstract: Computing posterior modes (e.g., maximum likelihood estimates) for models involving latent variables or missing data often involves complicated optimization procedures. By splitting this task into two simpler parts, however, EM-type algorithms often offer a simple solution. Although this approach has proven useful, in some settings even these simpler tasks are challenging. In particular, computations involving latent variables are typically difficult to simplify. Thus, in models such as hierarchical models with complicated latent variable structures, computationally intensive methods may be required for the expectation step of EM. This paper describes how nesting two or more EM algorithms can take advantage of closed form conditional expectations and lead to algorithms which converge faster, are straightforward to implement, and enjoy stable convergence properties. Methodology to monitor convergence of nested EM algorithms is developed using importance and bridge sampling. The strategy is applied to hierarchical probit and t regression models to derive algorithms which incorporate aspects of Monte-Carlo EM, PX-EM, and nesting in order to combine computational efficiency with easy implementation.

Key words and phrases: Bridge sampling, efficient data augmentation, Gibbs sampler, GLMM, hierarchical models, importance sampling, MCEM algorithm, MCMC, probit models, t-models, working parameters.

1. Introduction

The EM algorithm (Dempster, Laird and Rubin (1977)) is a popular method for computing maximum likelihood estimates, or more generally posterior modes, in the presence of missing data, or in models that can be formulated as such, latent-variable models for example. (Henceforth we refer to both missing data and latent variables as latent variables.) EM dichotomizes the often complex computational task of fitting a latent-variable model into two relatively simple steps, the expectation or E-step and the maximization or M-step. This, along with stable convergence properties, are arguably the reasons for the EM algorithm's popularity in practice. This said, however, the EM algorithm has often been criticized for its slow convergence when the fraction of missing information is large. There have been many strategies developed in the literature to speed up the EM algorithm (see McLachlan and Krishnan (1997) for a general discussion).

One set of methods, which directly reduces the fraction of missing information by either transforming the missing data or adjusting the missing data model, is especially attractive, in that it maintains the stability and simplicity of EM while improving its rate of convergence (Fessler and Hero (1994), Meng and van Dyk (1997, 1998), Liu, Rubin and Wu (1998), van Dyk (2000a)). Here we develop a new strategy for reducing the fraction of missing information. It involves nesting two EM algorithms and will be useful in EM algorithms with computationally expensive E-steps stemming from complex latent structures.

The Monte-Carlo EM algorithm (Wie and Tanner (1990)) accomplishes the E-step via Monte-Carlo integration. Although Monte-Carlo EM has been a popular method in practice, with many and diverse applications, it can be very slow to converge. One especially expensive computational strategy for implementing the algorithm (McCulloch (1994), Meng and Schilling (1996), Chan and Kuk (1997), Levine and Casella (1999), and others) is to use a Gibbs sampler (or a Metropolis algorithm, McCulloch (1997)) to obtain random draws for Monte-Carlo integration within each E-step. The nesting strategy developed here allows us to take advantage of closed form conditional expectations, which are relatively quick to compute, in order to improve the rate of convergence of the EM algorithm in such situations.

Our nesting strategy can be motivated by considering the intrinsic link between EM-type algorithms and the Gibbs sampler. Suppose, for example, we wish to sample from $p(\theta)$, where $\theta = (\theta_1, \theta_2, \theta_3)$ by using a Gibbs sampler which samples from each of $p(\theta_1|\theta_2, \theta_3)$, $p(\theta_2|\theta_1, \theta_3)$, and $p(\theta_3|\theta_1, \theta_2)$ in turn. Suppose also that sampling from $p(\theta_1|\theta_2, \theta_3)$ is expensive relative to sampling from the other two complete conditionals. In this case, it may be beneficial to sample once from $p(\theta_1|\theta_2, \theta_3)$ and then to sample from $p(\theta_2|\theta_1, \theta_3)$ and $p(\theta_3|\theta_1, \theta_2)$ K times each in turn. The benefit stems from the fact that if K is large, we are essentially sampling from $p(\theta_1|\theta_2, \theta_3)$ and $p(\theta_2, \theta_3|\theta_1)$, that is, we are using a blocked Gibbs sampler (see Liu, Wong and Kong (1995) and Roberts and Sahu (1997) for discussion on the advantage of blocking). Thus, the partially blocked Gibbs sampler will be useful when the advantage of blocking outweighs the cost of sampling from $p(\theta_2, \theta_3|\theta_1)$ via a nested Gibbs sampler (e.g., when θ_2 and θ_3 are highly correlated given θ_1). This strategy might be helpful when some of the complete conditionals are particularly difficult to sample (e.g., van Dyk et al. (2000)).

In the context of the EM algorithm, we can implement a similar strategy when the latent variables naturally divide into two or more parts, by taking advantage of the fact that the EM algorithm, which treats only one piece of the latent structure as missing by integrating over the rest, is faster to converge.

Although this algorithm will typically not have a closed form M-step, the maximization can be accomplished by a second, typically closed-form, EM algorithm that treats the remaining latent structure as missing. The resulting algorithm has an improved rate of convergence but, because of the nesting, each iteration will require more time. If the computational complexity of the E-step is relegated to the outer loop, this trade-off can go in favor of the nesting strategy when considering the actual computing time required.

The paper is organized as follows. In Section 2, we introduce the formal structure of the nested EM algorithm after a brief review of the EM algorithm. Properties, including monotone convergence in posterior or likelihood and faster convergence than the standard algorithm, will also be provided. Section 3 supplies advice on monitoring convergence using importance and bridge sampling, and on other practical issues. Two generalized linear mixed models are used in Section 4 to illustrate the current methodology and computational gain, and also to illustrate the combination of nesting, PX-EM (Liu, Rubin and Wu (1998)), and the Monte-Carlo EM algorithm. Concluding remarks appear in Section 5.

2. The EM and Nested EM Algorithms

2.1. Review of the EM algorithm

The EM algorithm and its various extensions are based upon a data augmentation scheme, Y_{aug}, defined such that the observed-data $Y_{\text{obs}} = \mathcal{M}(Y_{\text{aug}})$ for some many-to-one mapping $\mathcal{M}$. Starting with an initial estimate, $\theta^{(0)}$, of the parameter, the familiar two-step EM procedure iteratively computes θ^*, a mode of the log posterior, $\ell(\theta|Y_{\text{obs}}) = \log_p(\theta|Y_{\text{obs}})$ over $\theta \in \Theta$. The E-step of the $(t+1)$st iteration computes the conditional expectation of the augmented-data log posterior, $Q(\theta|\theta^{(t)}) = \mathrm{E}\left(\ell(\theta|Y_{\text{aug}})|Y_{\text{obs}}, \theta^{(t)}\right)$, where $\ell(\theta|Y_{\text{aug}}) = \log p(\theta|Y_{\text{aug}})$; the M-step then sets $\theta^{(t+1)}$ equal to $\arg\max_\theta\ Q(\theta|\theta^{(t)})$. It can be shown that this simple procedure increases the loglikelihood at each iteration and converges to a critical point of $\ell(\theta|Y_{\text{obs}})$, typically a mode in practice. The theoretical speed of convergence of the algorithm is determined by the matrix "fraction of missing information" or matrix rate of convergence (Dempster, Laird and Rubin (1977), Meng and Rubin (1994)), given by $DM^{EM} = I - I_{\text{obs}}I_{\text{aug}}^{-1}$, where $I_{\text{aug}}(\theta^*) = -D^{20}Q(\theta|\theta^*)\Big|_{\theta=\theta^*}$ is the expected augmented Fisher information matrix and I_{obs} is the observed Fisher information matrix. (Here D^{20} denotes the second partial derivative with respect to the first argument, etc., and we often denote $I_{\text{aug}}(\theta^*)$ by I_{aug}.) Under mild regularity conditions we define the global rate of convergence of EM as $\rho(DM^{EM}) = \lim_{t\to\infty} ||\theta^{(t+1)} - \theta^*||/||\theta^{(t)} - \theta^*||$, where $\rho(M)$ is the spectral radius of M.

2.2. The nested EM algorithm

When $Q(\theta|\theta^{(t)})$ cannot be computed in closed form the Monte-Carlo EM algorithm can be useful since $Q(\theta|\theta^{(t)})$ can always be approximated via Monte-Carlo integration if we can draw from $p(Y_{\text{aug}}|Y_{\text{obs}},\theta^{(t)})$ and evaluate $\ell(\theta|Y_{\text{aug}})$ many times. When the augmented data has complex structure, we may not be able to draw from $p(Y_{\text{aug}}|Y_{\text{obs}},\theta)$ directly. Suppose, for example, Y_{aug} can be subdivided into two (or more) parts, $Y_{\text{aug}} = (Y_{\text{obs}}, Y_{\text{mis}\,1}, Y_{\text{mis}\,2})$, such that $p(Y_{\text{mis}\,1}|Y_{\text{obs}}, Y_{\text{mis}\,2},\theta)$ and $p(Y_{\text{mis}\,2}|Y_{\text{obs}}, Y_{\text{mis}\,1},\theta)$ are both easy to sample directly, but $p(Y_{\text{aug}}|Y_{\text{obs}},\theta)$ is not. In this case we can run a Gibbs sampler to obtain draws from $p(Y_{\text{aug}}|Y_{\text{obs}},\theta)$ in order to implement the Monte-Carlo EM algorithm. Of course an EM algorithm with a Gibbs sampler within each iteration can be rather expensive computationally. The nested EM algorithm is designed to improve the performance of EM in this particular setting or, more generally, when the E-step is computationally expensive but the expected log posterior can be computed relatively quickly conditional on part of the augmented data.

For clarity, we will briefly examine the hierarchical probit regression model which will be discussed in detail in Section 4.1. Suppose

$$\omega_i = X_i\psi_i + e_i \quad e_i \sim N_{n_i}(0, I) \quad \psi_i \sim N_p(\mu, T) \quad \text{for} \quad i = 1,\ldots,m, \tag{2.1}$$

where $\omega_i = (\omega_{i1},\ldots,\omega_{in_i})^\top$ is an $(n_i \times 1)$ vector of censored responses for which we only observe $Y_{ij} = \text{sign}(\omega_{ij})$ for each i and j, X_i is an $(n_i \times p)$ completely observed matrix of covariates, and $\psi = (\psi_1,\ldots,\psi_m)$ are m unobserved $(p \times 1)$ random effects. Given $\theta = (\mu, T)$ and $\omega = (\omega_1,\ldots,\omega_m)$, ψ follows a multivariate normal distribution just as in the standard Gaussian hierarchical model. (Details appear in Section 4.1.) Likewise given θ, ψ, and the observed data $Y = (Y_{ij}, i = 1,\ldots,m, j = 1,\ldots,n_i)$, ω follows a truncated normal distribution. Although the joint distribution of the latent variables given Y and θ is not of a standard family, a Gibbs sampler can be used to obtain draws of the latent variables in order to run a Monte-Carlo EM algorithm. To make the most of this computationally expensive E-step, the nested EM algorithm will fix the augmented-data sufficient statistics involving, for example, ω and run several EM iterations conditional on these values since the E-step for this inner EM algorithm, based on the normal distribution $p(\psi|\omega, Y,\theta)$, is much cheaper.

To formalize this in the general setting, we introduce two nested data-augmentation schemes $Y_{\text{aug}\,1}$ and $Y_{\text{aug}\,2}$ such that $Y_{\text{obs}} = \mathcal{M}_1(Y_{\text{aug}\,1})$ and $Y_{\text{aug}\,1} = \mathcal{M}_2(Y_{\text{aug}\,2})$, for two many-to-one mappings $\mathcal{M}_1$ and $\mathcal{M}_2$. Likewise, we define two expected augmented-data log posteriors, $Q_1(\theta|\theta_0) = \text{E}\,(\ell(\theta|Y_{\text{aug}\,1})|Y_{\text{obs}},\theta_0)$ and $Q_2(\theta|\theta_0) = \text{E}\,(\ell(\theta|Y_{\text{aug}\,2})|Y_{\text{obs}},\theta_0)$, as well as the expectation of the corresponding function that treats the smaller data augmentation $Y_{\text{aug}\,1}$ as observed data, $Q_{21}(\theta|\theta_{01},\theta_{02}) = \text{E}\,(\text{E}\,(\ell(\theta|Y_{\text{aug}\,2})|Y_{\text{aug}\,1},\theta_{01})|Y_{\text{obs}},\theta_{02})$. (Note that

$Q_{21}(\theta|\theta_0,\theta_0) = Q_2(\theta|\theta_0)$, Q_1 and Q_2 are functions on $\Theta\times\Theta$ and Q_{21} is a function on $\Theta\times\Theta\times\Theta$). The tth iteration of the nested EM iteration repeats the following cycle K times, for $k = 1,\ldots,K$.

Cycle k for $k = 1,\ldots,K$:

E-step: Compute $Q_{21}(\theta|\theta^{(t+\frac{k-1}{K})},\theta^{(t)}) = \mathrm{E}\left(\mathrm{E}\left(\ell(\theta|Y_{\mathrm{aug}\,2})|Y_{\mathrm{aug}\,1},\theta^{(t+\frac{k-1}{K})}\right)|\right.$ $\left.Y_{\mathrm{obs}},\theta^{(t)}\right)$;

M-step: Set $\theta^{(t+\frac{k}{K})} = \arg\max_\theta Q_{21}(\theta|\theta^{(t+\frac{k-1}{K})},\theta^{(t)})$.

Upon completion of the Kth cycle, we set $\theta^{(t+1)} = \theta^{(t+\frac{K}{K})}$. Computational gain occurs when the outer expectation in the E-step only needs to be computed in the first cycle of each iteration. In particular, if we construct $Y_{\mathrm{aug}\,1}$ so that $\mathrm{E}\left(\ell(\theta|Y_{\mathrm{aug}\,2})|Y_{\mathrm{aug}\,1},\theta_{01}\right)$ is linear in $Y_{\mathrm{aug}\,1}$, we need only compute $Y_{\mathrm{aug}\,1}^{(t+1)} = \mathrm{E}\left(Y_{\mathrm{aug}\,1}|Y_{\mathrm{obs}},\theta^{(t)}\right)$ once per iteration and then we run K EM cycles with $(Y_{\mathrm{obs}}, Y_{\mathrm{aug}\,1}^{(t+1)})$ treated as observed data. We expect computational gain from this strategy when computing $Y_{\mathrm{aug}\,1}^{(t+1)}$ is expensive relative to the rest of the E-step and the M-step.

In the hierarchical probit model, we set $Y_{\mathrm{aug}\,1} = \{Y,\omega\}$ and $Y_{\mathrm{aug}\,2} = \{Y,\omega,\psi\}$ and use the Gibbs sampler to approximate $\mathrm{E}\left(\omega_i|Y_{\mathrm{obs}},\theta^{(t)}\right)$ and $\mathrm{E}\left(\omega_i\omega_i^\top|Y_{\mathrm{obs}},\theta^{(t)}\right)$ for $i = 1,\ldots,m$ in the first cycle. Since $\mathrm{E}\left(\ell(\theta|Y_{\mathrm{aug}\,2})|Y_{\mathrm{aug}\,1},\theta^{(t+\frac{k-1}{K})}\right)$ is linear in these statistics, $Q_{21}(\theta|\theta^{(t+\frac{k-1}{K})},\theta^{(t)})$ can easily be computed in closed form in subsequent cycles. Although ψ is drawn in the Gibbs sampler in the first cycle, we typically discard these samples and compute $Q_{21}(\theta|\theta^{(t)},\theta^{(t)})$ in the same manner as in the later cycles. That is, we compute $Q_{21}(\theta|\theta^{(t)},\theta^{(t)})$ using an iterated expectation as in the E-step above, rather than taking advantage of the fact that $Q_{21}(\theta|\theta^{(t)},\theta^{(t)}) = Q_2(\theta|\theta^{(t)})$. This both streamlines the code and takes advantage of Rao-Blackwellization.

2.3. Convergence of the nested EM algorithm

The nested EM algorithm enjoys the important convergence properties of EM including monotone convergence in posterior or likelihood. For the theoretical results, we assume the E-step is computed exactly.

Theorem 1. *Suppose $\{\theta^{(t)}, t \geq 0\}$ is a sequence in the parameter space computed with the nested EM algorithm, then $\ell(\theta^{(t+1)}) \geq \ell(\theta^{(t)})$ for each $t \geq 0$.*

Proof. It is sufficient to show that $Q_1(\theta^{(t+1)}|\theta^{(t)}) \geq Q_1(\theta^{(t)}|\theta^{(t)})$ (see Dempster, Laird and Rubin (1977, Theorem 1)). Thus, by construction of the algorithm, we need only show that $Q_1(\theta^{(t+\frac{k}{K})}|\theta^{(t)}) \geq Q_1(\theta^{(t+\frac{k-1}{K})}|\theta^{(t)})$ for $k = 1,\ldots,K$. To

obtain this, we note

$$Q_1(\theta|\theta^{(t)}) = Q_{21}(\theta|\theta', \theta^{(t)}) - \int\int \log p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta) p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta') \cdot p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t)}) dY_{\text{mis}\,2} dY_{\text{mis}\,1} + c,$$

for any θ' where c is a term not depending on θ. (By $\int \cdot dY_{\text{mis}\,1}$ we mean the integral with respect to $Y_{\text{aug}\,1}$ over the region $\{Y_{\text{aug}\,1} : \mathcal{M}_1(Y_{\text{aug}\,1}) = Y_{\text{obs}}\}$, likewise by $\int \cdot dY_{\text{mis}\,2}$ we mean the integral with respect to $Y_{\text{aug}\,2}$ over the region $\{Y_{\text{aug}\,2} : \mathcal{M}_2(Y_{\text{aug}\,2}) = Y_{\text{aug}\,1}\}$.) We use the above expression to evaluate the difference

$$\begin{aligned}
& Q_1(\theta^{(t+\frac{k}{K})}|\theta^{(t)}) - Q_1(\theta^{(t+\frac{k-1}{K})}|\theta^{(t)}) \\
&= Q_{21}(\theta^{(t+\frac{k}{K})}|\theta^{(t+\frac{k-1}{K})}, \theta^{(t)}) - Q_{21}(\theta^{(t+\frac{k-1}{K})}|\theta^{(t+\frac{k-1}{K})}, \theta^{(t)}) \qquad (2.2) \\
&\quad - \int\int \log\left(\frac{p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta^{(t+\frac{k}{K})})}{p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta^{(t+\frac{k-1}{K})})}\right) p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta^{(t+\frac{k-1}{K})}) \qquad (2.3) \\
&\qquad \cdot p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t)}) dY_{\text{mis}\,2} dY_{\text{mis}\,1}. \qquad (2.4)
\end{aligned}$$

The difference in (2.2) is positive by construction of the algorithm; that the inner integral in (2.3)-(2.4) is negative for each $Y_{\text{aug}\,1}$ follows from the Jensen inequality.

The next result asserts that the limit points of $\{\theta^{(t)}, t \geq 0\}$ are contained in $\{\Theta \in \Theta_0 : \frac{\partial}{\partial\theta}\ell(\theta|Y_{\text{obs}}) = 0\}$, where Θ_0 is the interior of Θ. The proof of this result, which assumes several standard regularity conditions used by Wu (1983), is omitted, but can be found in van Dyk (1998).

Theorem 2. *Assuming Wu's conditions* (6) - (10) *and that* $Q_{21}(\psi|\phi_1, \phi_2)$ *is continuous in all three of its arguments, all limit points of a nested EM sequence* $\{\theta^{(t)}, t \geq 0\}$ *are critical points of* $\ell(\theta|Y_{\text{obs}})$.

In order to evaluate the performance of the nested EM algorithm relative to the standard EM implementation which uses $Y_{\text{aug}\,2}$ as augmented data, we derive the matrix rate of convergence. We define $I_{\text{aug}\,i} = -D^{20}Q_i(\theta|\theta^*)|_{\theta=\theta^*}$ for $i = 1, 2$ and, using standard EM rate calculations, obtain three matrix rates of convergence. Noting that the first cycle of a nested EM algorithm corresponds to an iteration of an EM algorithm using data augmentation $Y_{\text{aug}\,2}$, we have $(\theta^{(t+\frac{1}{k})} - \theta^*) \approx (\theta^{(t)} - \theta^*)DM_2$ with $DM_2 = I - I_{\text{obs}}I^{-1}_{\text{aug}\,2}$. (Here $\approx$ indicates approximate equality for large t.) Second, the cycles within iteration $(t+1)$ form an EM algorithm which, under the regularity conditions below, converges to $\theta^*_{t+1} = \arg\max_\theta Q_1(\theta|\theta^{(t)})$, so that $(\theta^{(t+\frac{k}{K})} - \theta^*_{t+1}) \approx (\theta^{(t+\frac{k-1}{K})} - \theta^*_{t+1})DM_{21}(\theta^{(t)}, \theta^*_{t+1})$ with $DM_{21}(\theta^{(t)}, \theta^*_{t+1}) = I - I_{\text{aug}\,1}(\theta^{(t)})I^{-1}_{\text{aug}\,2}(\theta^*_{t+1})$. Finally, by the definition of θ^*_{t+1}, $(\theta^*_{t+1} - \theta^*) \approx (\theta^{(t)} - \theta^*)DM_1$ with $DM_1 = I - I_{\text{obs}}I^{-1}_{\text{aug}\,1}$.

Define $DM^{NEM}(\theta^{(t)}) = \partial\theta_j^{(t+1)}/\partial\theta_i^{(t)}$ and note asymptotically $\partial\theta_j^{(t+\frac{1}{K})}/\,\partial\theta_i^{(t)}|_{\theta=\theta^*}$ $= DM_2$ and using the definition of the inner EM mapping for $k = 2, \ldots, K$,

$$\frac{\partial\theta_j^{(t+\frac{k}{K})}}{\partial\theta_i^{(t)}} = \frac{\partial\theta^{(t+\frac{k-1}{K})}}{\partial\theta_i^{(t)}} \cdot \left(\frac{\partial\theta_j^{(t+\frac{k}{K})}}{\partial\theta^{(t+\frac{k-1}{K})}}\right)^\top + \frac{\partial\theta_{t+1}^*}{\partial\theta_i^{(t)}} \cdot \left(\frac{\partial\theta_j^{(t+\frac{k}{K})}}{\partial\theta_{t+1}^*}\right)^\top.$$

Solving recursively and evaluating at $\theta^{(t)} = \theta^*$, we obtain the following result.

Theorem 3. *If in addition to the conditions of Theorem* 2, $Q_1(\theta|\theta'), Q_2(\theta|\theta')$ *and* $Q_{12}(\theta|\theta', \theta'')$ *are log convex with modes contained in* Θ_0 *for each* $\theta', \theta'' \in \Theta$, *the nested EM algorithm has matrix rate of convergence given by*

$$DM^{NEM} = (DM_2 - DM_1)(DM_{21}(\theta^*, \theta^*))^{K-1} + DM_1. \tag{2.5}$$

The regularity conditions can be relaxed; see Theorem 3 in Dempster, Laird and Rubin (1977).

Since $Y_{\text{aug}\,1} = \mathcal{M}_2(Y_{\text{aug}\,2})$, we expect the matrix rate DM_1 to be preferable to DM_2 (since $I_{\text{aug}\,2} > I_{\text{aug}\,1}$, using a positive semidefinite ordering). Notice that as K increases, the matrix rate DM^{NEM} converges to DM_1. In fact, for $K = 1$ and $K = \infty$ the matrix rate is exactly DM_2 and DM_1 respectively. If $I_{\text{aug}\,1} = I_{\text{aug}\,2}$, that is, the two data augmentations are identical in terms of their information for the parameters, $DM^{NEM} = DM_2$, that is, the nested EM algorithm is identical to the EM algorithm but unnecessarily repeats the M-step K times. On the other hand if $I_{\text{aug}\,1} = I_{\text{obs}}$, $DM^{NEM} = (DM_2)^K$, the nested algorithm will be K times faster (in terms of the number of iterations), but each iteration will take K times longer. Thus, as illustrated in the examples in Section 4, the nested EM algorithm will be advantageous when additional cycles are relatively cheap computationally, but $I_{\text{aug}\,2}$ is much larger than $I_{\text{aug}\,1}$.

The final result shows how the global rate of convergence depends on K.

Theorem 4. *Under the conditions of Theorem* 3, *the global rate of convergence of the nested EM algorithm decreases* (*i.e., improves*) *with* K.

Proof. The global rate is defined to be the spectral radius of the matrix rate. This can be written as $DM_K^{NEM} = I - I_{\text{obs}}B_K$, where $B_K = [(I_{\text{aug}\,2}^{-1} - I_{\text{aug}\,1}^{-1})$ $(I_{\text{mis}\,2}I_{\text{aug}\,2}^{-1})^{K-1} + I_{\text{aug}\,1}^{-1}]$ with $I_{\text{mis}\,2} = I_{\text{aug}\,2} - I_{\text{aug}\,1}$. It suffices to show that $B_1 > 0$ and for each $K > 1$, $B_K > B_{K-1}$, since the global rate of convergence of the nested EM algorithm is the largest eigenvalue of $I - I_{\text{obs}}^{1/2}B_K I_{\text{obs}}^{1/2}$. Now, $B_1 = I_{\text{aug}\,2}^{-1} > 0$ and

$$\begin{aligned} B_K - B_{K-1} &= (I_{\text{aug}\,1}^{-1} - I_{\text{aug}\,2}^{-1})(I_{\text{aug}\,1}I_{\text{aug}\,2}^{-1})(I_{\text{mis}\,2}I_{\text{aug}\,2}^{-1})^{K-2} \\ &= I_{\text{aug}\,2}^{-1}(I_{\text{aug}\,2} - I_{\text{aug}\,1})I_{\text{aug}\,2}^{-1}(I_{\text{mis}\,2}I_{\text{aug}\,2}^{-1})^{K-2} \end{aligned}$$

$$= I_{\text{aug}\,2}^{-1}(I_{\text{mis}\,2}I_{\text{aug}\,2}^{-1})^{K-1},$$

which is positive definite. This completes the proof.

3. Implementation of the Nested EM Algorithm

3.1. Detecting convergence

In models that are formulated in terms of complex latent structures, convergence of mode-finding algorithms can be difficult to ascertain since $\ell(\theta|Y_{\text{obs}})$ may be difficult to evaluate. In particular, the integral

$$p(\theta|Y_{\text{obs}}) \propto p(\theta) \int\int p(Y_{\text{aug}\,2}|\theta) dY_{\text{mis}\,2} dY_{\text{mis}\,1} \tag{3.1}$$

may have no analytical solution. When a Monte-Carlo E-step is used, the situation is even more complicated since, unless the Monte-Carlo sample size L_t at iteration t grows with t, $\{\theta^{(t)}, t \geq 0\}$ will not converge, necessitating extra numerical or graphical methods to determine convergence of the algorithm (see Wie and Tanner (1990), McCulloch (1994, 1997), and others).

In order to evaluate $\ell(\theta^{(t+1)}|Y_{\text{obs}}) - \ell(\theta^{(t)}|Y_{\text{obs}}) = \log p(\theta^{(t+1)}) - \log p(\theta^{(t)}) + \log \delta^{(t+1)}$, where $\delta^{(t+1)} = p(Y_{\text{obs}}|\theta^{(t+1)})/p(Y_{\text{obs}}|\theta^{(t)})$, a common strategy takes advantage of the fact that $p(Y_{\text{aug}\,2}|\theta)$ and $p(Y_{\text{aug}\,1}|\theta)$ are often easy to evaluate. (Here, $p(Y_{\text{aug}\,1}|\theta)$ is defined similarly to (3.1), but the integral is of smaller dimension.) The well-known importance sampling estimate of $\delta^{(t+1)}$ is (Chan and Ledolter (1995), and others)

$$\hat{\delta}_i^{(t+1)} = \frac{1}{L_t} \sum_{l=1}^{L_t} \frac{p(Y_{\text{aug}\,i}^{(l)}|\theta^{(t+1)})}{p(Y_{\text{aug}\,i}^{(l)}|\theta^{(t)})} \quad \text{for} \quad i = 1 \text{ or } 2, \tag{3.2}$$

where $\{Y_{\text{aug}\,i}^{(l)}, l = 1, \ldots, L_t\}$ is a sample from $p(Y_{\text{aug}\,i}|Y_{\text{obs}}, \theta^{(t)})$. When using a Monte-Carlo E-step this technique is especially attractive as the necessary draws are a byproduct of the E-step.

Intuitively, $\hat{\delta}_1^{(t+1)}$ should be better behaved than $\hat{\delta}_2^{(t+1)}$ since the Monte-Carlo integration is of lower dimension. Formally, we can prove that with independent Monte-Carlo draws, asymptotically $RE^2(\hat{\delta}_1^{(t+1)}) \leq RE^2(\hat{\delta}_2^{(t+1)})$ where $RE^2(\hat{\delta}^{(t+1)}) = \text{E}\,(\hat{\delta}^{(t+1)} - \delta^{(t+1)})^2/(\delta^{(t+1)})^2$ is the relative mean-square error. In practice, we may not have independent draws, but the result gives guidance as to the best choice and confirms our intuition. The proof involves showing that the chi-square distance between $p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t+1)})$ and $p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t)})$ is dominated by that between $p(Y_{\text{aug}\,2}|Y_{\text{obs}}, \theta^{(t+1)})$ and $p(Y_{\text{aug}\,2}|Y_{\text{obs}}, \theta^{(t)})$, but is omitted since it is similar to the argument outlined below for the typically superior bridge sampling estimates.

An improvement over the importance sampling approximations to $\delta^{(t+1)}$ can be obtained by using samples from both $p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t)})$ and $p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t+1)})$, rather than just the former. Bridge sampling (Meng and Wong (1996)) accomplishes this via the identity

$$\begin{aligned}&\frac{p(Y_{\text{obs}}|\theta)}{p(Y_{\text{obs}}|\theta')}\frac{\int p(Y_{\text{aug}\,i}|\theta')\phi(Y_{\text{aug}\,i})p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta)dY_{\text{mis}}}{\int p(Y_{\text{aug}\,i}|\theta)\phi(Y_{\text{aug}\,i})p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta')dY_{\text{mis}}}\\&=\frac{\int p(Y_{\text{aug}\,i}|\theta')\phi(Y_{\text{aug}\,i})p(Y_{\text{aug}\,i}|\theta)dY_{\text{mis}}}{\int p(Y_{\text{aug}\,i}|\theta)\phi(Y_{\text{aug}\,i})p(Y_{\text{aug}\,i}|\theta')dY_{\text{mis}}}\end{aligned}\tag{3.3}$$

for any $\phi(Y_{\text{aug}\,i})$ such that $0<|\int\phi(Y_{\text{aug}\,i})p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t)})p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t+1)})dY_{\text{mis}}|<\infty$. Since (3.3) equals one, $\delta^{(t+1)}$ can be approximated by setting $\theta=\theta^{(t)}$ and $\theta'=\theta^{(t+1)}$, with

$$\tilde{\delta}_i^{(t+1)}=\frac{\frac{1}{L_{t+1}}\sum_{l=1}^{L_{t+1}}p(Y_{\text{aug}\,i}^{(l)}|\theta^{(t+1)})\phi(Y_{\text{aug}\,i}^{(l)})}{\frac{1}{L_t}\sum_{l=1}^{L_t}p(\tilde{Y}_{\text{aug}\,i}^{(l)}|\theta^{(t)})\phi(\tilde{Y}_{\text{aug}\,i}^{(l)})}\quad\text{for}\quad i=1\,\text{or}\,2,$$

where $\{\tilde{Y}_{\text{aug}\,i}^{(l)},l=1,\ldots,L_{t+1}\}$ is a sample from $p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t+1)})$. Although in the general bridge sampling setting, an optimal ϕ (in terms of minimizing $RE^2(\tilde{\delta}^{(t+1)})$) is not available without numerical iteration, Meng and Wong (1996) suggest $\phi(Y_{\text{aug}\,i})=1/\sqrt{p(Y_{\text{aug}\,i}|\theta^{(t)})p(Y_{\text{aug}\,i}|\theta^{(t+1)})}$, yielding

$$\tilde{\delta}_i^{(t+1)}=\log\frac{\frac{1}{L_t}\sum_{l=1}^{L_t}\sqrt{p(Y_{\text{aug}\,i}^{(l)}|\theta^{(t+1)})/p(Y_{\text{aug}\,i}^{(l)}|\theta^{(t)})}}{\frac{1}{L_{t+1}}\sum_{l=1}^{L_{t+1}}\sqrt{p(\tilde{Y}_{\text{aug}\,i}^{(l)}|\theta^{(t)})/p(\tilde{Y}_{\text{aug}\,i}^{(l)}|\theta^{(t+1)})}}\quad\text{for}\quad i=1\,\text{or}\,2,\tag{3.4}$$

since it stabilizes the importance ratios and tends to perform well in practice.

Assuming $p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(s)})$ has the same support for each s, the asymptotic relative error for $\tilde{\delta}_i^{(t+1)}$ with independent samples is given by

$$\begin{aligned}&RE^2(\tilde{\delta}_i^{(t+1)})\\&=\frac{L_t+L_{t+1}}{L_tL_{t+1}}\left(\left[1-\frac{1}{2}H^2\left(p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t)}),p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t+1)})\right)\right]^{-2}-1\right),\end{aligned}$$

which is an increasing function of $0\le H^2(p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t)}),\,p(Y_{\text{aug}\,i}|Y_{\text{obs}},\theta^{(t+1)})\le 2$, the square of the Hellinger distance,

$$H(p_1(\omega),p_2(\omega))=\left[\int\left(\sqrt{p_1(\omega)}-\sqrt{p_2(\omega)}\right)^2d\omega\right]^{1/2}=\left[2-2\int\sqrt{p_1(\omega)p_2(\omega)}d\omega\right]^{1/2}.\tag{3.5}$$

That $\tilde{\delta}_1^{(t+1)}$ is a better estimate of $\delta^{(t+1)}$ than is $\tilde{\delta}_2^{(t+1)}$ is evident since

$$\int\int\left(p(Y_{\text{aug}\,2}|Y_{\text{obs}},\theta^{(t)})p(Y_{\text{aug}\,2}|Y_{\text{obs}},\theta^{(t+1)})\right)^{1/2}dY_{\text{mis}\,2}dY_{\text{mis}\,1}$$

$$= \int \left(p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t)})p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t+1)}\right)^{1/2}$$
$$\cdot \int \left(p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta^{(t)})p(Y_{\text{aug}\,2}|Y_{\text{aug}\,1}, \theta^{(t+1)}\right)^{1/2} dY_{\text{mis}\,2} dY_{\text{mis}\,1}$$
$$\leq \int \left(p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t)})p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t+1)}\right)^{1/2} dY_{\text{mis}\,1}.$$

Thus by (3.5), $H(p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t)}), p(Y_{\text{aug}\,1}|Y_{\text{obs}}, \theta^{(t+1)})) \leq H(p(Y_{\text{aug}\,2}|Y_{\text{obs}}, \theta^{(t)}), p(Y_{\text{aug}\,2}|Y_{\text{obs}}, \theta^{(t+1)}))$, and we have the following result.

Theorem 5. *If for $i = 1$ or 2 $p(Y_{\text{aug}\,i}|Y_{\text{obs}}, \theta^{(s)})$ has the same support for each $s \geq 0$, then (asymptotically),*

$$RE^2(\tilde{\delta}_1^{(t+1)}) \leq RE^2(\tilde{\delta}_2^{(t+1)})$$

(*assuming independent draws*).

Bridge sampling is illustrated and compared with importance sample in Section 4.1.

3.2. Choosing K and L_t

When using the nested EM algorithm with Monte-Carlo integration, there are two parameters that typically need to be set by the user: K, the number of cycles in each EM iteration, and L_t, the number of Monte-Carlo draws in the first E-step of each iteration. In choosing K, the goal is *not* to reach convergence of the inner EM algorithm (i.e., convergence to θ^*_{t+1}), but rather to make as much progress towards the (local) mode of $\ell(\theta|Y_{\text{obs}})$ as we can with small computational cost. Since the EM algorithm typically makes substantial progress in its early iterations, we expect the first few cycles to make significant progress towards the mode, but later cycles to be rather costly relative to their progress. Thus, we typically recommend a moderate value of K. Of course, K can vary between iterations, and $\ell(\theta^{(t+\frac{k}{K})}, Y_{\text{obs}})$ or some function of the parameter can be monitored to determine "convergence" of the inner EM algorithm. For example,

$$\log\left(\frac{p(\theta^{(t+\frac{k}{K})}|Y_{\text{obs}})}{p(\theta^{(t)}|Y_{\text{obs}})}\right) - \log\left(\frac{p(\theta^{(t+\frac{k-1}{K})}|Y_{\text{obs}})}{p(\theta^{(t)}|Y_{\text{obs}})}\right) \tag{3.6}$$

can be computed at each iteration using importance sampling. (Since we do not have draws from $p(Y_{\text{aug}}|Y_{\text{obs}}, \theta^{(t+\frac{k}{K})})$ and obtaining such draws would defeat the purpose of nested EM, we cannot use Bridge sampling.) There is no theoretical guarantee that (3.6) will be positive at each cycle; we only know that the posterior will increase at each iteration. In practice, however, we expect (3.6) to be positive.

When evaluating $\ell(\theta|Y_{\text{obs}})$ is computationally expensive, we recommend fixing K at some small value (say between 2 and 10). If the inner EM algorithm is

slow to converge, a large value of K is better, e.g., $K = 10$; if it is fast to converge a small value is fine, e.g., $K = 3$. In fact if the inner algorithm converges very fast, i.e., $I_{\text{aug}\,1} \approx I_{\text{aug}\,2}$, there is essentially no reduction in the fraction of missing information due to nesting and the standard EM algorithm will suffice, i.e., $K = 1$. Thus, if there is a choice as to how the latent variables are divided between $Y_{\text{aug}\,1}$ and $Y_{\text{aug}\,2}$, it is advantageous to make the information in $Y_{\text{aug}\,1}$ as "small" as possible in order to "decrease" DM_1. Since DM^{NEM} is a weighted average of DM_1 and DM_2 (fixed), this will enable us to further reduce the fraction of missing information for the nested EM algorithm. Since this will simultaneously increase the difference $I_{\text{aug}\,2} - I_{\text{aug}\,1}$ and thereby increase the fraction of missing information that controls DM_{21}, the rate of convergence of the inner EM algorithm, K will have to be increased to take advantage of this gain, see (2.5).

When implementing a Monte-Carlo E-step, it is necessary to choose L_t at each iteration. Clearly larger values will result in more exact but slower calculations. (Thus, larger values will be useful when verifying that $\ell(\theta|Y_{\text{obs}})$ is increasing at each iteration while debugging the code.) A typical strategy is to let L_t grow as a function of t to ensure both quick convergence at first and more precise calculations later. (When computing modes as a preliminary step in preparation for the Gibbs sampler and indeed for most inferential purposes, the position of the modes need not be computed with exacting precision.) Wei and Tanner (1990), for example, suggest starting with a moderate value and increasing it when graphical inspection shows that the parameter or a function thereof has stabilized. Chan and Ledolter (1995) suggest drawing several samples of size L_0 and running Gibbs samplers to estimate the Monte-Carlo variance which in turn can be used to compute the necessary Monte-Carlo sample size for a desired level of precision. An intermediate strategy, as suggested by McCulloch (1994), increases L_t by some small amount at each iteration, which ensures accuracy in the later iterations, is fast in the early iterations, and requires little intervention.

4. Examples

In this section, we describe two applications of the nested (Monte-Carlo) EM algorithm, fitting hierarchical probit and t models. See van Dyk (2000b) for an application to speetral analysis that does not involve a Monte-Carlo E-step.

4.1. Hierarchical probit model

Returning to the example introduced in Section 2.2, we outline a Monte-Carlo EM algorithm for hierarchical probit regression similar to algorithms developed by McCulloch (1994, 1997) and Chan and Kuk (1997) and describe how nesting can decrease the required computational time. For model (2.1)

we set $Y_{\text{aug}} = \{Y, \omega, \psi\}$ and implement a Gibbs sampler to obtain draws from $p(\omega, \psi|Y, \theta^{(t)})$, where $\theta = (\mu, T)$, for the Monte-Carlo evaluation of

$$Q(\theta|\theta^{(t)}) = -\frac{m}{2}\log|T| - \frac{1}{2}\sum_{i=1}^{m}\text{tr}\left(T^{-1}\text{E}\left[(\psi_i-\mu)(\psi_i-\mu)^\top|Y,\theta^{(t)}\right]\right)$$

in the E-step. Throughout we assume a flat prior on all parameters to compute the maximum likelihood estimates. In particular, we can obtain $\{\psi_i^{(l)}, l = 1,\ldots,L_t, i = 1,\ldots,m\}$ by iteratively drawing from two complete conditional distributions of the latent variables, first

$$\psi_i|\omega, Y, \theta, \sim N_p(\hat{\psi}_i(\omega_i), T - TX_i^\top W_i X_i T), \quad i = 1,\ldots,m,$$

where $\hat{\psi}_i(\omega_i) = \mu + TX_i^\top W_i(\omega_i - X_i\mu)$ with $W_i = [I_{n_i} + X_i^\top T X_I]^{-1}$ and second $p(\omega_i|\psi, Y, \theta) \propto N_{n_i}(X_i\psi_i, I_{n_i})$, where the normal distribution is truncated so that $\text{sign}(\omega_{ij}) = Y_{ij}$ for $j = 1,\ldots,n_i$. Finally, $Q(\theta|\theta^{(t)})$ can be evaluated by setting

$$\text{E}\left[\psi_i|Y,\theta^{(t)}\right] \approx \frac{1}{L_t}\sum_{l=1}^{L_t}\psi_i^{(l)} \qquad \text{and} \qquad \text{E}\left[\psi_i\psi_i^\top|Y,\theta^{(t)}\right] \approx \frac{1}{L_t}\sum_{l=1}^{L_t}\psi_i^{(l)}[\psi_i^{(l)}]^\top. \tag{4.1}$$

Here and below $\approx$ indicates the Monte-Carlo approximation. The M-step then maximizes $Q(\theta|\theta^{(t)})$ by setting

$$\mu^{(t+1)} = \frac{1}{m}\sum_{i=1}^{m}\text{E}\left[\psi_i|Y,\theta^{(t)}\right] \tag{4.2}$$

and

$$T^{(t+1)} = \frac{1}{m}\sum_{i=1}^{m}\left(\text{E}\left[\psi_i\psi_i^\top|Y,\theta^{(t)}\right] - \text{E}\left[\psi_i|Y,\theta^{(t)}\right]\text{E}\left[\psi_i|Y,\theta^{(t)}\right]^\top\right), \tag{4.3}$$

completing the tth iteration of the EM algorithm.

In order to reduce Monte-Carlo error, we can rewrite (4.1) in the Rao-Blackwellizied form

$$\text{E}\left[\psi_i|Y,\theta^{(t)}\right] = \text{E}\left[\hat{\psi}_i(\omega_i)|Y,\theta^{(t)}\right] \approx \hat{\psi}_i(\mathbf{S}_{1i}), \text{ where } \mathbf{S}_{1i} \equiv \frac{1}{L_t}\sum_{l=1}^{L_t}\omega_i^{(l)} \approx \text{E}\left[\omega_i|Y,\theta^{(t)}\right] \tag{4.4}$$

and

$$\text{E}\left[\psi_i\psi_i^\top|Y,\theta^{(t)}\right] \approx \text{E}\left[\hat{\psi}_i(\omega_i)[\hat{\psi}_i(\omega_i)]^\top|Y,\theta^{(t)}\right] + T - TX_i^\top W_i X_i T. \tag{4.5}$$

To evaluate the first term on the right hand side of (4.5), write

$$\begin{aligned}
&\text{E}\left[\hat{\psi}_i(\omega_i)[\hat{\psi}_i(\omega_i)]^\top|Y,\theta^{(t)}\right] \\
&\approx \left[(I-TX_i^\top W_i X_i)\mu\right]\left[(I-TX_i^\top W_i X_i)\mu\right]^\top + \left[(I-TX_i^\top W_i X_i)\mu\right]\left[TX_i^\top W_i\mathbf{S}_{1i}\right]^\top \\
&\quad + \left[TX_i^\top W_i\mathbf{S}_{1i}\right]\left[(I-TX_i^\top W_i X_i)\mu\right]^\top + TX_i^\top W_i\mathbf{S}_{2i}W_i X_i T,
\end{aligned}$$

where $\mathbf{S}_{2i} = \sum_{l=1}^{L_t} \omega_i^{(l)} [\omega_i^{(l)}]^\top / L_t \approx \mathrm{E}\,[\omega_i \omega_i^\top | Y, \theta^{(t)}]$. Although the advantage of the Rao-Blackwellizied estimate may be small when L_t is large, (4.4) and (4.5) highlight the fact that once we have evaluated $\mathbf{S}_{1i}$ and $\mathbf{S}_{2i}$ we can reevaluate $\mathrm{E}\,[\mathrm{E}\,[\ell(\theta|Y_{\mathrm{aug}})|Y, \omega, \theta']|Y, \theta^{(t)}]$ for any θ' without the expensive Monte-Carlo step. In particular, to implement the nested EM algorithm we set $Y_{\mathrm{aug}\,1} = \{Y, \omega\}$ and $Y_{\mathrm{aug}\,2} = \{Y, \psi, \omega\}$. In the first cycle of each iteration we first evaluate $(\mathbf{S}_{1i}, \mathbf{S}_{2i})$, for $i = 1, \ldots, m$ via the Gibbs sampler Monte-Carlo estimates. The E-step is completed by evaluating (4.4), (4.5) and the M-step consists of (4.2) and (4.3). In subsequent cycles of each iteration, only (4.4), (4.5), (4.2) and (4.3) are recomputed using the same $(\mathbf{S}_{1i}, \mathbf{S}_{2i})$ for $i = 1, \ldots, m$.

In order to further improve computational performance, we can introduce a working parameter to reduce the fraction of missing information (Meng and van Dyk (1997, 1999), Liu, Rubin and Wu (1998)). For example, we can rewrite $Y_{\mathrm{aug}\,2}$ as $\{Y, \sigma\psi, \sigma\omega\}$, where σ^2 is an unidentifiable working parameter and define $\zeta_i = \sigma\psi_i$ and $\varpi_i = \sigma\omega_i$ for $i = 1, \ldots, m$, $Y_{\mathrm{aug}\,1} = \{Y, \varpi\}$, and $Y_{\mathrm{aug}\,2} = \{Y, \varpi, \zeta\}$. (See van Dyk and Meng (1999) for other potential working parameters.) In order to evaluate

$$Q_2(\theta|\theta^{(t)}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{m} \mathrm{E}\left[(\varpi_i - X_i\zeta_i)^\top(\varpi_i - X_i\zeta_i)|Y, \theta^{(t)})\right]$$
$$-\frac{m}{2}\log|\sigma^2 T| - \frac{1}{2\sigma^2}\sum_{i=1}^{m} \mathrm{tr}\left(T^{-1}\mathrm{E}\left[(\zeta_i - \sigma\mu)(\zeta_i - \sigma\mu)^\top|Y, \theta^{(t)}\right]\right),$$

where $n = \sum_{i=1}^{m} n_i$ in the E-step, we again run a Gibbs sampler. Because the observed-data model does not depend on σ^2, we can set σ^2 to 1 at the beginning of each E-step, in which case $\omega_i = \varpi_i$ and $\psi_i = \zeta_i$ for each i. Thus, the introduction of the working parameters does not alter the Gibbs sampler. We compute $(\mathbf{S}_{1i}, \mathbf{S}_{2i})$ for each i exactly as before and approximate $\mathrm{E}\,[\zeta_i|Y, \theta^{(t)}]$ and $\mathrm{E}\,[\zeta_i\zeta_i^\top|Y, \theta^{(t)}]$ with the expressions given in (4.4) and (4.5). To evaluate the first term of $Q_2(\theta|\theta^{(t)})$, we also compute

$$\mathrm{E}\,[\varpi_i^\top X_i\zeta_i|Y, \theta^{(t)}] \approx \mathbf{S}_{1i}^\top X_i\mu + \mathrm{tr}\left(X_i T X_i^\top W_i \mathbf{S}_{2i}\right) - \mathbf{S}_{1i}^\top X_i T X_i^\top W_i X_i\mu \quad (4.6)$$

for $i = 1, \ldots, m$, which follows from $\mathrm{E}\,[\zeta_i|Y_{\mathrm{aug}\,1}, \theta]$ and the law of iterated expectations. This completes the E-step.

The M-step updates the parameters as follows:

$$[\sigma^2]^{(t+1)} = \frac{1}{n}\sum_{i=1}^{m}\left(\mathbf{S}_{2i} - 2\mathrm{E}\,[\varpi_i^\top X_i\zeta_i|Y, \theta^{(t)}] + \mathrm{tr}(X_i\mathrm{E}\,[\zeta_i\zeta_i^\top|Y, \theta^{(t)}]X_i^\top)\right), \quad (4.7)$$

$$[\sigma\mu]^{(t+1)} = \frac{1}{m}\sum_{i=1}^{m}\mathrm{E}\left[\zeta_i^\top|Y, \theta^{(t)}\right], \quad (4.8)$$

and

$$(\sigma^2 T)^{(t+1)} = \frac{1}{m}\sum_{i=1}^{m}\left(\mathrm{E}\left[\zeta_i\zeta_i^\top|Y,\theta^{(t)}\right] - \mathrm{E}\left[\zeta_i|Y,\theta^{(t)}\right]\mathrm{E}\left[\zeta_i|Y,\theta^{(t)}\right]^\top\right). \qquad (4.9)$$

Finally, by the invariance of maximum likelihood estimates, we have $\mu^{(t+1)} = [\sigma\mu]^{(t+1)}/\sqrt{[\sigma^2]^{(t+1)}}$ and $T^{(t+1)} = [\sigma^2 T]^{(t+1)}/[\sigma^2]^{(t+1)}$. Thus, in the E-step of the Monte-Carlo/Parameter-Expanded EM algorithm we first compute $(\mathbf{S}_{1i}, \mathbf{S}_{2i})$ for $i = 1, \ldots, m$ via the Monte-Carlo estimates and evaluate $\mathrm{E}[\zeta_i|Y,\theta^{(t)}]$, $\mathrm{E}[\zeta_i\zeta_i^\top|Y,\theta^{(t)}]$, and $\mathrm{E}[\varpi_i^\top X_i\zeta_i|Y,\theta^{(t)}]$ as in (4.4), (4.5), and (4.6) respectively. The iteration is completed by the M-step as given by (4.7) – (4.9) with the transformation to the original parameters. For the nested Monte-Carlo/ Parameter Expanded EM algorithm, the first cycle of each iteration is exactly the same as this EM iteration, but in subsequent cycles only (4.4) – (4.9) are recomputed, and this does not require the Gibbs sampler.

In order to illustrate the computational gain resulting from nesting, we fit (2.1) to an artificial data set, consisting of $m = 20$ groups with sizes n_i varying between six and ten, with a total of $n = 160$ observations. Each observation consists of a success/failure indicator and a single covariate. Model (2.1) is fit with this covariate and an intercept. Evaluating the effect of nesting on Monte-Carlo EM algorithms is made difficult by the many strategies possible for choosing the Monte-Carlo sample sizes L_t, which can greatly affect the relative performance of the algorithms. In order to give a general idea of the relative efficiencies, however, we implemented two strategies for choosing L_t. The first fixed L_t at 100, (McCulloch (1997) and Chan and Kuk (1997) fix L_t at values ranging from 50 to 1000) with the idea that in practice the accuracy of the approximation to the mode would have to be evaluated and the algorithm rerun with a larger value of L_t if the accuracy is not sufficient. The second strategy attempted to automate the procedure by setting $L_t = 50t$. We emphasize that we are not advocating either of these strategies for general implementation. It is impossible to give advice regarding the choice of L_t that is generally applicable and some trial and error will always be necessary. We believe, however, that these two strategies are typical of the type that are often useful in practice and illustrate well the benefit of nesting.

Figure 1 illustrates the progress of the EM and nested EM algorithms in terms of the loglikelihood evaluated at the iterations. The algorithms were run using the working parameter σ^2 and with $K = 3, 5, 7$, and 15. The upper panel illustrates that the algorithm that uses no nesting takes roughly two times longer to reach the vicinity of the mode. We also see that the choice of K is not critical, as each of the four other algorithms are roughly equivalent. The second panel highlights the computational gain of nesting by plotting the number of times

the unnested algorithm takes longer than each of the nested algorithms to first exceed a given value of the loglikelihood. The lines start out together on the far left because each algorithm was run with the same starting value. The oscillation on the far right is due to the oscillation in the loglikelihood. This is illustrated in the final panel which focuses on values of the loglikelihood near the mode and shows that a larger number of Monte-Carlo draws is required if more accurate

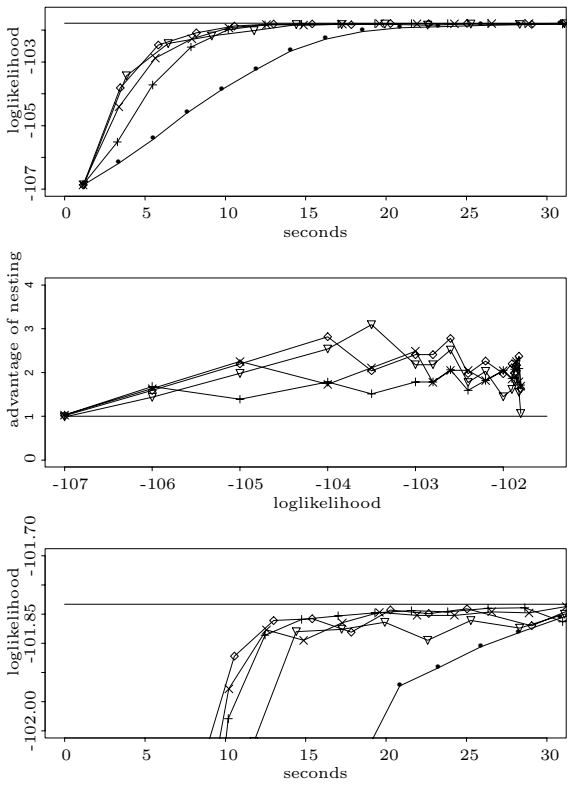

Figure 1. Fitting a Hierarchical Probit Model with 100 Monte-Carlo Draws per Iteration. The first plots shows the increase in loglikelihood (the modal value is represented by the horizontal line) for an EM algorithm (iterations indicated by dots) and several nested EM algorithms (iterations represented as follows: $K = 3$ by plus signs; $K = 5$ by $\times$ sign; $K = 7$ by boxes; and $K = 15$ by triangles). The second plot shows the number of times longer the EM algorithm took to first pass a given value of the loglikelihood (horizontal axis) than each of the nested algorithms. The third plot is a close up of the top of the first plot. The value of K does not seem to be critical and nesting tends to reduce the computational time by a factor of about two in this example.

results are necessary. Figure 2 is similar to Figure 1, except that the corresponding algorithms were run with $L_t = 50t$. This results in somewhat slower convergence, but a steadily increasing loglikelihood. Nesting again decreases the required computation time, this time by a factor of between three and four. In this case the smallest value of K, 3, tends to perform somewhat worse than the other values of K in the earlier iterations.

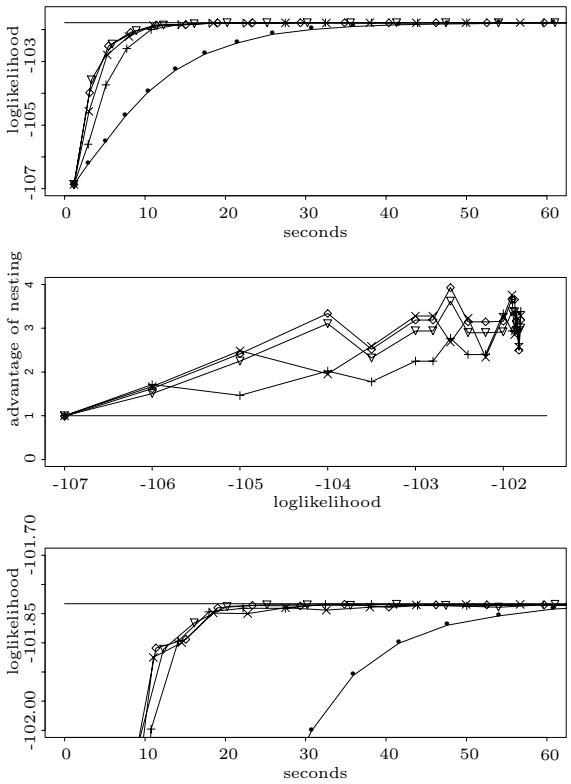

Figure 2. Fitting a Hierarchical Probit Model Increasing the Number of Monte-Carlo Draws by 50 at Each Iteration. The components of this figure are identical to those of Figure 1. Here, however, the algorithms are run with $50t$ Monte-Carlo draws at iteration t. By increasing the number of draws as the algorithms proceed, the maximizer is computed more accurately with less intervention by the user. Here nesting reduces computational time by a factor of three or four.

In these figures, the loglikelihood was computed via numerical integration. In (3.1) the integration over ω can be accomplished analytically, leaving $Y_{ij}|\psi_i, \theta \sim$ Binomial$(n = 1, p = \Phi(X_{ij}\psi_i))$ for $i = 1, \ldots, m$ and $j = 1, \ldots, n_i$, where $\Phi(\cdot)$ is

the standard normal cumulative density function and X_{ij} is the jth row of X_i. Thus, we can compute the likelihood via the p dimensional integral,

$$p(Y|\theta) \propto \prod_{i=1}^{m} \int \prod_{j=1}^{n_i} (\Phi(X_{ij}\psi_i))^{Y_{ij}} (1 - \Phi(X_{ij}\psi_i))^{(1-Y_{ij})} |T|^{-1/2} \cdot \exp\Big((\psi_i - \mu)^\top T^{-1}(\psi_i - \mu)/2\Big)\, d\psi_i. \tag{4.10}$$

In this example $p = 2$ and the integral is relatively easy to evaluate. For larger p, however, direct evaluation becomes difficult and the bridge sampling techniques described in Section 3.1 are useful.

Here, we use the numerical evaluation of (4.10) to evaluate the accuracy and relative accuracy of importance and bridge sampling in this problem. In both (3.2) and (3.4), theory suggests using $Y_{\text{aug}\,1} = \{Y, \omega\}$ in place of $Y_{\text{aug}\,2} = \{Y, \omega, \psi\}$ since the dimension of the numerical integration is reduced. We will use an alternative augmented data set $\{Y, \psi\}$, which could be defined to be $Y_{\text{aug}\,1}$ in the derivation of a different set of nested EM algorithms, since the resulting numerical integration is of smaller dimension, p rather than n_i. That is, we compute (3.2) and (3.4) with $Y_{\text{aug}\,1}$ replaced by $\{Y, \psi\}$. The results appear in Figure 3 for L_t fixed at both 100 and 5000 with $K = 1$ and $K = 7$. The plots show the absolute value of the difference between the step sizes as computed with either bridge or importance sampling and (4.10), as a function of the iteration number. (Note the difference in scale used in the several plots.) The dashed line represents importance sampling, the dotted line bridge sampling, and the solid line the actual signal as computed with (4.10). As expected, it is clear from the plots that bridge sampling tends to outperform importance sampling. For the small steps at the end of the iteration, the Monte-Carlo error swamps the signal, especially with only 100 Monte-Carlo draws. We emphasize, however, that these techniques are useful for code-debugging, since in software development, bridge sampling can be used with large Monte-Carlo sample sizes to verify that the loglikelihood (or log posterior) is increasing at each iteration.

4.2. Hierarchical t model

As a second example, we consider a Gaussian hierarchical model,

$$Y_i = X_i\beta_i + e_i\ , \qquad e_i \sim N\Big(0, \frac{\sigma^2}{v_i} I_{n_i}\Big) \quad \text{for} \quad i = 1, \ldots, m,$$

with Y_i an $(n_i \times 1)$ response vector and X_i an $(n_i \times p)$ matrix of covariates. In this model we assume a distribution not only on the regression coefficients,

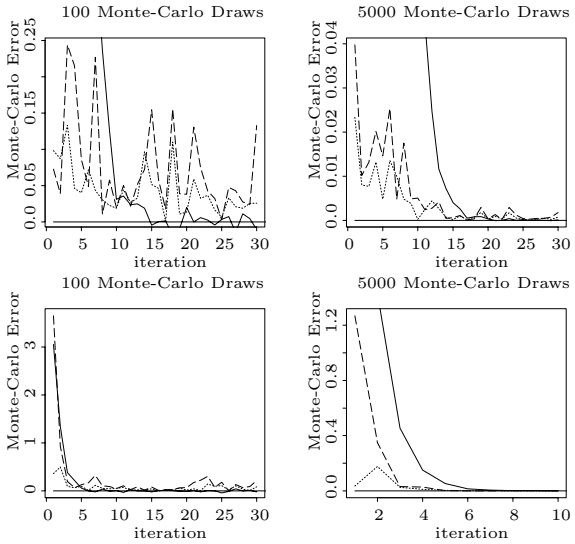

Figure 3. Approximating Step Sizes Using Importance and Bridge Sampling. The four plots compare importance sampling with bridge sampling for computing the step size in loglikelihood of each iteration. The actual step size as computed by numerical integration is given by the solid line, the other two lines record the difference between the solid line and what was reported by importance sampling (dashed line) and bridge sampling (dotted line). The plots in the left column correspond to 100 Monte-Carlo draws; those in the right column correspond to 5000 draws. The first row represents the EM algorithm, the second the nested EM algorithm with $K = 7$. Notice that for the early iterations (especially with a larger number of Monte-Carlo draws) both importance sampling and bridge sampling do a reasonable job relative to the size of the signal given by the solid line. Bridge sampling, however, tends to out perform importance sampling. Since the nested algorithm takes bigger steps, the error in both importance and bridge sampling is larger than with the standard EM algorithm.

$\beta_i \sim N(\mu, T)$ as in the common random-effects model, but also on the variances, through the latent variable, $v_i \sim \chi^2_\nu/\nu$, where ν is typically fixed and represents the variability among the group variances. We refer to this model as a hierarchical t-model since the conditional distribution of Y_i given β_i is multivariate t. In order to derive an EM algorithm, we define $Y_{\text{aug}} = \{(Y_i, \beta_i, v_i), i = 1, \ldots, m\}$, but note that $Q(\theta|\theta^{(t)})$, with $\theta = (\sigma, \mu, T)$, cannot be evaluated in closed form and must be approximated via Monte-Carlo integration. In particular, given θ we can run a Gibbs sampler using the complete conditional distributions of the latent variables which are given by

$$\beta_i|v, Y, \theta \sim N_p(\hat{\beta}_i(W_i(v_i)), T - TX_i^\top W_i(v_i)X_iT), \quad i = 1, \ldots, m,$$

where $Y = (Y_1, \ldots, Y_m)$, $v = (v_1, \ldots, v_m)$,

$$\hat{\beta}_i(W_i(v_i)) = \mu + TX_i^\top W_i(v_i)(Y_i - X_i\mu) \quad \text{with} \quad W_i(v_i) = \left[\frac{\sigma^2}{v_i} I_{n_i} + X_i^\top T X_i\right]^{-1}$$

and

$$v_i|\beta, Y, \theta \sim Gamma\left(\frac{\nu+1}{2}, \frac{1}{2(\nu + d_i^2)}\right),$$

when $\beta = (\beta_1, \ldots, \beta_m)$ and $d_i^2 = (Y_i - X_i\beta_i)^\top (Y_i - X_i\beta_i)/\sigma^2$. (The Gamma distribution is parameterized so that $\mathrm{E}[\nu_i|\beta, Y, \theta] = (\nu+1)/(\nu + d_i^2)$.) The M-step maximizes $Q(\theta|\theta^{(t)})$ by setting

$$\mu^{(t+1)} = \frac{1}{m}\sum_{i=1}^m \mathrm{E}\left[\beta_i|Y, \theta^{(t)}\right], \tag{4.11}$$

$$T^{(t+1)} = \frac{1}{m}\sum_{i=1}^m \left(\mathrm{E}\left[\beta_i\beta_i^\top|Y, \theta^{(t)}\right] - \mathrm{E}\left[\beta_i|Y, \theta^{(t)}\right]\mathrm{E}\left[\beta_i|Y, \theta^{(t)}\right]^\top\right), \tag{4.12}$$

and

$$\begin{aligned}[\sigma^2]^{(t+1)} &= \frac{1}{n}\sum_{i=1}^m \mathrm{E}\left[v_i(Y_i - X_i\beta_i)^\top(Y_i - X_i\beta_i)|Y, \theta^{(t)}\right] \\ &= \frac{1}{n}\sum_{i=1}^m \left(\mathrm{E}\left[v_i|Y, \theta^{(t)}\right] Y_i^\top Y_i - 2Y_i^\top X_i \mathrm{E}\left[v_i\beta_i|Y, \theta^{(t)}\right]\right. \qquad (4.13) \\ &\quad \left. + \operatorname{tr}\left(X_i \mathrm{E}\left[v_i\beta_i\beta_i^\top|Y, \theta^{(t)}\right] X_i^\top\right)\right). \qquad (4.14)\end{aligned}$$

The five different expectations in (4.11) and (4.14) are approximated using a Monte-Carlo E-step with the Gibbs sampler described above. Again we use Rao-Blackwellizied forms:

$$\mathrm{E}\left[\beta_i|Y, \theta^{(t)}\right] = \mathrm{E}\left[\hat{\beta}_i(W_i(v_i))|Y, \theta^{(t)}\right] \approx \hat{\beta}_i(\mathbf{S}_{1i}) \quad \text{where} \quad \mathbf{S}_{1i} = \frac{1}{L_t}\sum_{l=1}^{L_t} W_i(v_i^{(l)}), \tag{4.15}$$

with $\{v_i^{(1)}, \ldots, v_i^{(L_t)}\}$ being the L_t draws of v_i at iteration t (for $i = 1, \ldots, m$),

$$\mathrm{E}\left[\beta_i\beta_i^\top|Y, \theta^{(t)}\right] = \mathrm{E}\left[\hat{\beta}_i(W_i(v_i))[\hat{\beta}_i(W_i(v_i))]^\top|Y, \theta^{(t)}\right] + T - TX_i^\top \mathbf{S}_{1i} X_i T, \tag{4.16}$$

where $\mathrm{E}[\hat{\beta}_i(W_i(v_i))[\hat{\beta}_i(W_i(v_i))]^\top|Y, \theta^{(t)}] \approx \hat{\beta}_i(\mathbf{S}_{1i})\mu^\top + \mu(\hat{\beta}_i(\mathbf{S}_{1i}) - \mu)^\top + TX_i^\top \mathbf{S}_{2i} X_i T$ with $\mathbf{S}_{2i}$ given by $\frac{1}{L_t}\sum_{l=1}^{L_t} W_i(v_i^{(l)})(Y_i - X_i\mu)(Y_i - X_i\mu)^\top W_i(v_i^{(l)})$,

$$\mathrm{E}[v_i|Y, \theta^{(t)}] \approx \mathbf{S}_{3i} \quad \text{where} \quad \mathbf{S}_{3i} = \frac{1}{L_t}\sum_{l=1}^{L_t} v_i^{(l)}, \tag{4.17}$$

$$\mathrm{E}\left[v_i\beta_i|Y,\theta^{(t)}\right] = \mathrm{E}\left[v_i\hat{\beta}_i(W_i(v_i))|Y,\theta^{(t)}\right] \approx \mu\mathbf{S}_{3i} + TX_i^\top\mathbf{S}_{4i}(Y - X_i\mu), \quad \text{where}$$

$$\mathbf{S}_{4i} = \frac{1}{L_t}\sum_{l=1}^{L_t} v_i^{(l)}W_i(v_i^{(l)}), \tag{4.18}$$

and

$$\begin{aligned}\mathrm{E}\left[v_i\beta_i\beta_i^\top|Y,\theta^{(t)}\right] =& \mathrm{E}\left[v_i\left(\hat{\beta}_i(W_i(v_i))[\hat{\beta}_i(W_i(v_i))]^\top + T - TX_i^\top W_i(v_i)X_iT\right)|Y,\theta^{(t)}\right]\\ \approx& \mathbf{S}_{3i}\mu\mu^\top + TX_i^\top\mathbf{S}_{4i}(Y - X_i\mu)\mu^\top + \left[TX_i^\top\mathbf{S}_{4i}(Y - X_i\mu)\mu^\top\right]^\top\\ &+ TX_i^\top\mathbf{S}_{5i}X_iT + \mathbf{S}_{3i}T - TX_i^\top\mathbf{S}_{4i}X_iT,\end{aligned} \tag{4.19}$$

where $\mathbf{S}_{5i} = \frac{1}{L_t}\sum_{l=1}^{L_t} v_i^{(l)}W_i(v_i^{(l)})(Y_i - X_i\mu)(Y_i - X_i\mu)^\top W_i(v_i^{(l)})$. Thus, one iteration of the EM algorithm first computes $\{\mathbf{S}_{ji}, j = 1,\ldots,5, i = 1,\ldots,m\}$ using L_t Monte-Carlo draws of v, then uses these statistics to compute the expectations given in (4.15) – (4.19), and finally updates the parameters via (4.11) – (4.14). In the nested EM algorithm we set $Y_{\text{aug}\,1} = \{Y, v\}$ and $Y_{\text{aug}\,2} = \{Y, v, \beta\}$ so that $\{(v_i^{(1)},\ldots,v_i^{(L_t)}), i = 1,\ldots,m\}$ only needs to be drawn in the first cycle of each iteration. In subsequent cycles, the E-step consists of reevaluating $\{\mathbf{S}_{ji}, j = 1,2,4,5, i = 1,\ldots,m\}$ and (4.15) – (4.19) using the updated parameter values, and the M-step is given by (4.11) and (4.14). Because $\mathrm{E}\left(\ell(\theta|Y_{\text{aug}\,2})|Y_{\text{aug}\,1},\theta_0\right)$ is not linear in v, the Monte-Carlo estimates $\mathbf{S}_{ji}$ for $j \neq 3$ must be computed at each cycle of the algorithm. Despite this, nesting significantly reduces the computational requirement of the algorithm because the Gibbs sampler is only run during the first cycle of each iteration.

In order to further increase computational efficiency, we introduce a working parameter into the marginal distribution of v_i. In particular, we replace $Y_{\text{aug}\,1} = \{Y, \beta, v\}$ with $Y_{\text{aug}\,1} = \{Y, \beta, u\}$, where $u_i = \alpha v_i$ for $i = 1,\ldots,m$. Computationally, the effect of this change is simply that (4.13) – (4.14) is replaced by

$$\begin{aligned}[\sigma^2]^{(t+1)} = \frac{1}{\sum_{i=1}^m n_i\mathbf{S}_{3i}}\sum_{i=1}^m\Big(&\mathrm{E}\left[v_i|Y,\theta^{(t)}\right]Y_i^\top Y_i - 2Y_i^\top X_i\mathrm{E}\left[v_i\beta_i|Y,\theta^{(t)}\right]\\ &+ \mathrm{tr}(X_i\mathrm{E}\left[v_i\beta_i\beta_i^\top|Y,\theta^{(t)}\right]X_i^\top)\Big).\end{aligned}$$

That is, the division by n is replaced by division by the "sum of the weights". See Meng and van Dyk (1997) for discussion of this substitution in the non-hierarchical t-model.

Figure 4 illustrates the relative computational cost of the EM and the nested EM algorithms using $K = 3, 5, 7$, and 15. The hierarchical t model with $\nu = 10$ was fit to data from an orthodontic study (Potthoff and Roy (1964)). The

distance between the pteryomaxillary fissure and the center of the pituitary was measured on 16 boys at age 8, 10, 12, and 14. Here Y_i are the four measurements for boy i, and the covariates consist of an intercept term and age. In Figure 4, L_t was set to $20t$. Although there is some variability as a function of K in the amount of improvement, nesting reduces the computation time by a factor of between three (when K is small) and ten (when K is large).

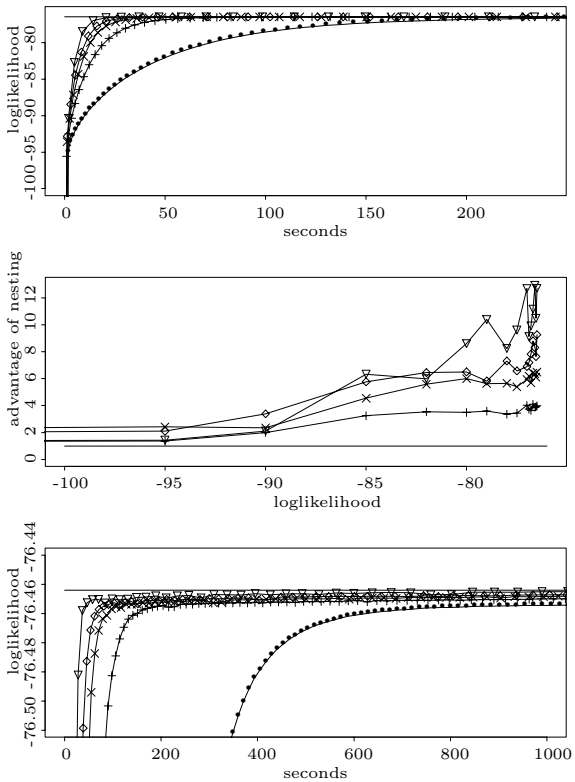

Figure 4. Fitting a Hierarchical t-Model Increasing the Number of Monte-Carlo Draws by 20 at Each Iteration. The components of this figure are identical to those of Figure 1. Again we increase the number of draws at each iteration to stabilize the convergence of the algorithms. The nested algorithms, espescially with large K, show substantial improvement over the standard algorithm.

5. Discussion

The EM algorithm is a useful tool for computing maximum likelihood estimates and posterior modes for models involving latent variables or missing data. Even when MCMC methods are used to map out the posterior or likelihood

more completely, the location of the (multiple) modes is invaluable for determining starting vales and the EM computer code typically forms the backbone of the computer code for the Gibbs sampler (so programming effort is kept to a minimum). In models formulated in terms of multiple latent variables, standard optimization techniques such as Newton-Raphson type algorithms not only suffer from stability problems when $\ell(\theta|Y)$ is far from quadratic, but also can be difficult to implement when numerical integration is required to evaluate $\ell(\theta|Y)$. In these settings the (Monte-Carlo) EM algorithm offers a stable albeit slow solution, requiring little programming that is not required by the Gibbs sampler itself. By nesting we can both maintain the stability of EM and increase the computational efficiency while requiring essentially no extra effort by those who implement the algorithm.

Acknowledgements

The research was supported in part by the NSF grant DMS 97-05156 and by the U.S. Census Bureau. The author thanks an anonymous referee for many helpful comments.

References

Chan, J. S. K. and Kuk, A. Y. C. (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **53**, 86-97.

Chan, K. S. and Ledolter, J. (1995). Monte Carlo EM estimation for time series models involving counts. *J. Amer. Statist. Assoc.* **90**, 242-252.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood estimation from incomplete-data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39**, 1-38.

Fessler, J. A. and Hero, A. O. (1994). Space-alternating generalized expectation-maximization algorithm. *IEEE Tran. on Signal Processing* **42**, 2664-2677.

Levine, R. A. and Casella, G. (1999). Implementations to the Monte Carlo EM algorithm. Submitted to *J. Comput. Graphical Statist.*

Liu, C., Rubin, D. B. and Wu, Y. N. (1998). Parameter expansion for EM acceleration: the PX-EM algorithm. *Biometrika* **85**, 755-770.

Liu, J. S., Wong, W. H. and Kong, A. (1995). Covariance structure and convergence rate of the Gibbs sampler with various scans. *J. Roy. Statist. Soc. Ser. B* **57**, 157-169.

McCulloch, C. E. (1994). Maximum likelihood variance components estimation for binary data. *J. Amer. Statist. Assoc.* **89**, 330-335.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *J. Amer. Statist. Assoc.* **92**, 162-170.

McLachlan, G. J. and Krishnan, T. (1997). *The EM Algorithm and Extensions.* John Wiley & Sons, New York.

Meng, X. L. and Schilling, S. (1996). Fitting full-information item factor models and an empirical investigation of bridge sampling. *J. Amer. Statist. Assoc.* **91**, 1254-1267.

Meng, X. L. and Rubin, D. B. (1994). On the global and componentwise rates of convergence of the EM algorithm. *Linear Algebra and Its Applications (Special issue honoring Ingram Olkin)* **199**, 413-425.

Meng, X. L. and van Dyk, D. A. (1997). The EM algorithm – an old folk song sung to a fast new tune (with discussion). *J. Roy. Statist. Soc. Ser. B* **59**, 511-567.

Meng, X. L. and van Dyk, D. A. (1998). Fast EM implementations for mixed-effects models. *J. Roy. Statist. Soc. Ser. B* **60**, 559-578.

Meng, X. L. and van Dyk, D. A. (1999). Seeking efficient data augmentation schemes via conditional and marginal augmentation. *Biometrika* **86**, 301-320.

Meng, X. L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statist. Sinica* **6**, 831-860.

Potthoff, R. F. and Roy, S. N. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika* **51**, 313-326.

Roberts, G. O. and Sahu, S. K. (1997). Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. Roy. Statist. Soc. Ser. B.* **59**, 291-317.

van Dyk, D. A. (1998). Nesting EM algorithms for computational efficiency. Unpublished Technical Report.

van Dyk, D. A. (2000a). Fitting mixed-effects models using efficient EM-type algorithms. *J. Comput. Graphical Statist.* To appear.

van Dyk, D. A. (2000b). Fast new EM-type algorithms with applications in astrophysics. Submitted to *The Astrophysical J.*

van Dyk, D. A., Conners, A., Kashyap, V. and Siemiginowska, A. (2000). Analysis of energy spectra with low photor counts via Bayesian posterior simulation. Unpublished Technical Report.

van Dyk, D. A. and Meng, X. L. (1999). The art of data augmentation. Submitted to *Ann. Statist.*

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm. *J. Amer. Statist. Assoc.* **85**, 699-704.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95-103.

Department of Statistics, Harvard University, Science Center, One Oxford Street, Cambridge, MA 02138, U.S.A.

E-mail: vandyk@stat.harvard.edu

(Received April 1998; accepted March 1999)