J. R. Statist. Soc. B (1997)
59, *No.* 3, *pp.* 511–567

The EM Algorithm — an Old Folk-song Sung to a Fast New Tune

By XIAO-LI MENG† and DAVID VAN DYK

University of Chicago, USA *Harvard University, Cambridge, USA*

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, December 11th, 1996,* Professor P. J. Green *in the Chair*]

SUMMARY

Celebrating the 20th anniversary of the presentation of the paper by Dempster, Laird and Rubin which popularized the EM algorithm, we investigate, after a brief historical account, strategies that aim to make the EM algorithm converge faster while maintaining its simplicity and stability (e.g. automatic monotone convergence in likelihood). First we introduce the idea of a 'working parameter' to facilitate the search for efficient data augmentation schemes and thus fast EM implementations. Second, summarizing various recent extensions of the EM algorithm, we formulate a general alternating expectation–conditional maximization algorithm AECM that couples flexible data augmentation schemes with model reduction schemes to achieve efficient computations. We illustrate these methods using multivariate t-models with known or unknown degrees of freedom and Poisson models for image reconstruction. We show, through both empirical and theoretical evidence, the potential for a dramatic reduction in computational time with little increase in human effort. We also discuss the intrinsic connection between EM-type algorithms and the Gibbs sampler, and the possibility of using the techniques presented here to speed up the latter. The main conclusion of the paper is that, with the help of statistical considerations, it is possible to construct algorithms that are simple, stable *and* fast.

Keywords: DATA AUGMENTATION; EXPECTATION–CONDITIONAL MAXIMIZATION ALGORITHM; EXPECTATION–CONDITIONAL MAXIMIZATION EITHER ALGORITHM; GIBBS SAMPLER; INCOMPLETE DATA; MARKOV CHAIN MONTE CARLO METHOD; MISSING DATA; MODEL REDUCTION; MULTIVARIATE t-DISTRIBUTIONS; POISSON MODEL; POSITRON EMISSION TOMOGRAPHY; RATE OF CONVERGENCE; SAGE ALGORITHM

1. PROLOGUE: HISTORY AND POPULARITY

1.1. *Who First Developed the EM Algorithm?*

With the ever growing popularity of the EM algorithm, especially with its various deterministic and stochastic extensions (e.g. the data augmentation algorithm of Tanner and Wong (1987)), those of us who do research in this area find ourselves being asked more frequently the question who first developed the EM algorithm? Although it is easy for us to direct the inquirer to Dempster *et al.* (1977) where the term EM appeared for the first time, the question is really not easy to answer. In fact, the issue of the origin of the EM method was raised by several discussants of Dempster *et al.* (1977). For example, Hartley opened his contribution with

> 'I felt like the old minstrel who has been singing his song for 18 years and now finds, with considerable satisfaction, that his folklore is the theme of an overpowering symphony'.

†*Address for correspondence*: Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, IL 60637-1514, USA.
E-mail: meng@galton.uchicago.edu

 0035–9246/97/59511

Hartley's 'folk-song' analogy is indeed appropriate for describing the development of this powerful method. Just as a folk-song typically evolves many years before its tune is well recognized, various EM-type methods or ideas which precede Dempster *et al.* (1977), and in fact precede Hartley (1958) by many years, can be found in the literature. For instance, the earliest piece of the EM score traced by Dempster *et al.* (1977) is McKendrick (1926). If we are willing to make a broader connection, then a key identity underlying the EM algorithm can be traced back, as with many other popular statistical methods (e.g. the bootstrap), to the work of Fisher (i.e. Fisher (1925)), as pointed out by Efron in his discussion of Dempster *et al.* (1977).

The folk-song analogy is also accurate in the sense that it signifies the collective effort in developing the EM algorithm. Indeed, a couple of dozen individuals were credited by Dempster *et al.* (1977) for contributing to one degree or another, some with new verses and some with remakes. Among them, Baum *et al.* (1970) is perhaps the most sophisticated—we still cannot sing it without first warming up with the version of Dempster *et al.* (1977). This is not a criticism of Baum *et al.* (1970), who might have been required by their publisher to adopt such a compact version, but merely a remark attempting to explain why their version, which had the key notes as did Dempster *et al.* (1977), did not become the hit that Dempster *et al.* (1977) did seven years later. Combining Baum *et al.* (1970) with Sundberg (1974, 1976), which were based on the author's thesis at Stockholm University (Sundberg, 1972), perhaps would have caught more attention. Sundberg not only provided an easily accessible rendition of the theory underlying the EM algorithm when the complete data are from an exponential family (where the algorithm is most useful) but also illustrated the iterative method with several examples. What was missing in Sundberg's version was an explicit result on the monotone convergence in likelihood, a celebrated feature of the EM algorithm, which was proved in Baum *et al.* (1970). As a further note on the difficulties in answering the question of the origin of the EM algorithm, Sundberg (1976) acknowledged that his key 'iteration mapping', which corresponds to the EM mapping defined by Dempster *et al.* (1977), was suggested by A. Martin-Löf in a personal communication.

Although we shall perhaps never be able to find out who really sang the first musical note of the EM algorithm, we all agree that it was Dempster *et al.* (1977) who brought it into the all-time top 10 of statistics (see Stigler (1994)). They made (at least) two contributions that popularized the song. First, they gave it an informative title identifying the key stanzas—the expectation step and the maximization step. Second, they demonstrated how it can be sung at many different occasions, some of which had not previously been thought to be related to the EM algorithm (e.g. viewing latent variables as missing data). Since then, we all have sung or heard it being sung many times, sometimes with abusive or even unbearable tones.

1.2. *EM: A Bibliographical Review with Missing Articles*

This section shares its title with Meng and Pedlow (1992), who conducted a bibliographical search of EM-related papers (using their definition) in statistical and non-statistical literatures. This turned out to be an essentially hopeless task, not only because of the ever increasing number of EM papers (see Fig. 1), but also because many papers applied the EM method without citing any reference or listing it in the keywords, much like when we use a Taylor series expansion or the Newton–Raphson algorithm.

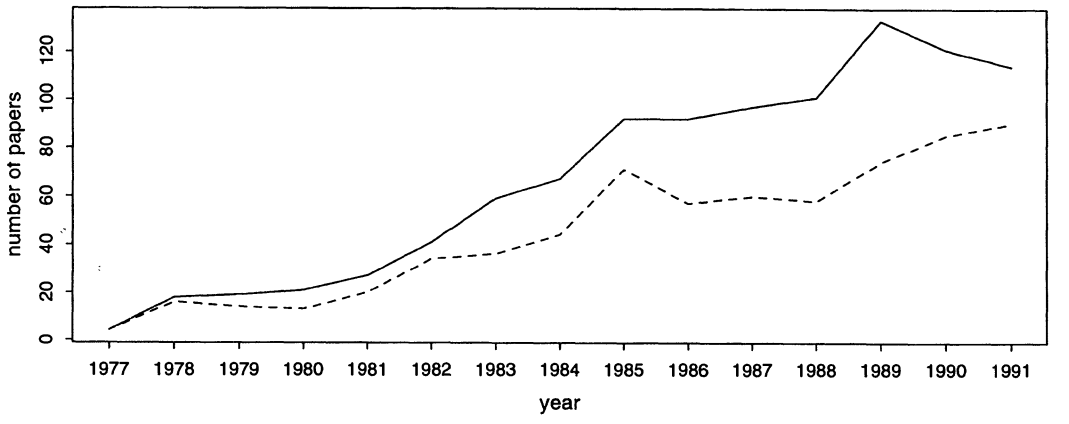

Fig. 1. Annual number of EM-related articles (———) and the number of papers that cited Dempster *et al.* (1977) (- - - -): the data are from Meng and Pedlow's preliminary EM bibliography (available on request) and were reported in Meng and Pedlow (1992); they noted that the decrease after 1989 should not be interpreted as a real decline, because the number of 'missing articles' increases over the years, and because there is a 'reporting delay' in the seven citation sources that they used for the search (for example, the 1991 *Current Index to Statistics* was not available at the time that the bibliography was prepared)

Nevertheless, even with many missing articles, Meng and Pedlow (1992) found over 1000 EM-related articles appearing in almost 300 journals, about 85% of which are non-statistical, the majority of which many statisticians never consult. Meng and Pedlow (1992) also reported the 12 journals that have published the most EM-related papers from 1977 to 1991, according to their preliminary EM bibliography (Table 1). As Meng and Pedlow (1992) noted, it is not surprising that the list includes the most popular statistical journals, with the one exception of *The Annals of Statistics*, indicating the applied nature of the EM algorithm. The appearance of the *Journal of Dairy Science* in the fourth place, even before Series B of the *Journal of the Royal Statistical Society*, was unexpected and was mainly due to a series of articles on modelling animal breeding data via variance components models.

TABLE 1
Top 12 from Meng and Pedlow's (1992) preliminary EM bibliography

Journal	*No. of papers*
1. *Journal of the American Statistical Association*	82
2. *Biometrics*	68
3. *Psychometrika*	38
4. *Journal of Dairy Science*	29
5. *Journal of the Royal Statistical Society*, Series B	27
6. *Communications in Statistics A: Theory and Methods*	26
7. *Biometrika*	25
8. *IEEE Transactions on Signal Processing*	24
9. *IEEE Transactions on Nuclear Science*	21
10. *Applied Statistics*	21
11. *Technometrics*	20
12. *IEEE Transactions on Medical Imaging*	19
Total	400

There is no attempt here to update Meng and Pedlow's (1992) findings or to review the background of the algorithm. Instead, we refer readers to Vardi and Lee (1993), Lange (1995a, b) and the special theme topic in *Statistica Sinica* (volume 5, number 1, 1995) entitled 'EM and related algorithms' and the references therein for recent work on the application and theory of EM-type algorithms. Our purpose here is to show promising ways of tackling the most often voiced complaint about the algorithm, namely its slow convergence. The popularity of the EM algorithm stems from its simplicity and stability, both of which save human time. Our new tune centres on making it fast, measured in computational time, while *maintaining its simplicity and stability*. Our main methodological contribution (Section 2) is the introduction of the 'working parameter' approach to search for efficient data augmentation schemes to construct fast EM-type algorithms. Our main theoretical contribution (Section 3) is the formulation of the alternating expectation–conditional maximization (AECM) algorithm, which unifies several recent extensions of the EM algorithm that effectively combine data augmentation with model reduction. We present two examples, the t-model and the Poisson model for image reconstruction, to illustrate our methods and to show the potentially dramatic reduction in computing time with little increase in human effort. A third important application to the mixed effects model is presented in Meng and van Dyk (1997). We conclude with a discussion of the potential application of these methods to stochastic extensions of EM-type algorithms such as the Gibbs sampler.

2. AUGMENTING DATA EFFICIENTLY TO SPEED UP EM ALGORITHM

2.1. *Speeding up EM Algorithm with Little Sacrifice*

Although the principal reasons for the popularity of the EM algorithm are its easy implementation and stable convergence, various attempts have been made to speed it up since it can converge quite slowly in some applications. Methods proposed to speed up the algorithm include the use of Aitkin acceleration (e.g. Dempster *et al.* (1976) and Louis (1982)), combining it with Newton–Raphson-type algorithms (e.g. Lange (1995b)) or with conjugate gradient methods (e.g. Jamshidian and Jennrich (1993)), etc. A practically undesirable feature of these accelerations is that the savings in computer time are achieved typically at the expense of a larger human investment for general users. This is because these methods require not only more numerically complex implementations but also more careful monitoring (partially due to the lack of automatic monotone convergence), and even with such care the algorithms may not converge properly (e.g. Lansky and Casella (1990)). However, there is a way of improving the speed of the EM algorithm without much sacrifice of its simplicity or stability. Since the rate of convergence of the algorithm is determined by the fraction of missing information, the data augmentation scheme (i.e. defining the so-called complete data space) that one uses for constructing the augmented data likelihood (or posterior) determines the speed of the algorithm. It has been well recognized since Dempster *et al.* (1977) that, by augmenting less, one can develop a faster algorithm, but a common trade-off is that the resulting M-step and/or E-step may be more difficult to implement. If the M-step and E-step resulting from less augmentation are equally simple (or slightly less simple but the gain in speed is relatively substantial), then there is little reason not to use the faster EM algorithm. This, for example, motivated the expectation–conditional maximization either (ECME) algorithm (Liu

and Rubin, 1994) and the space alternating generalized EM (SAGE) algorithm (Fessler and Hero, 1994), both of which will be discussed in Section 3.

In this section we present an approach that uses this idea for accelerating the EM algorithm by searching for an efficient data augmentation scheme. By 'efficient' we mean less augmentation while *maintaining the simplicity and stability* of the EM algorithm. Our key idea is to introduce a working parameter, which is not in the underlying problem, to index a class of possible data augmentation schemes to facilitate our search. We illustrate our approach using multivariate t-models with known degrees of freedom. In Section 3 we shall use this approach in conjunction with the AECM algorithm to construct an efficient algorithm for t-models with unknown degrees of freedom and for image reconstruction under the Poisson model.

2.2. *An Optimal EM Algorithm for t-model with Known Degrees of Freedom*

The multivariate (including univariate) t-distribution is a common model for statistical analysis, especially for robust estimation (e.g. Lange *et al.* (1989)). Here we let $t_p(\mu, \Sigma, \nu)$ denote a p-dimensional t-variable with known degrees of freedom ν, and the density

$$f_\nu(x|\mu, \Sigma) \propto |\Sigma|^{-1/2}\{\nu + (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\}^{-(\nu+p)/2}, \qquad x \in \mathbb{R}^p.$$

Fitting this model to a data set $Y_{\text{obs}} = (y_1, \ldots, y_n)$ typically requires maximizing the likelihood function $\Pi_i f_\nu(y_i|\mu, \Sigma)$, which is known to have no general closed form solution. The standard EM algorithm for this problem uses a data augmentation scheme based on the following well-known representation:

$$t_p \equiv t_p(\mu, \Sigma, \nu) = \mu + \Sigma^{1/2}Z/\sqrt{q}, \qquad Z \sim N_p(0, I_p), \qquad q \sim \chi^2_\nu/\nu \qquad Z \perp q, \tag{2.1}$$

with I_p the p-dimensional identity matrix and '$\perp$' indicating independence. Now *assume* that y_i, $i = 1, \ldots, n$, are independent and identically distributed realizations of this t_p. Since t_p follows $N_p(\mu, \Sigma/q)$ conditionally on q, if we further assume that the q_i, $i = 1, \ldots, n$, were observed, i.e. $Y_{\text{aug}} = \{(y_i, q_i), i = 1, \ldots, n\}$ are our augmented data, finding the maximum likelihood estimate (MLE) of $\theta \equiv (\mu, \Sigma)$ follows directly from the well-known weighted least squares procedure (see equations (2.3) and (2.4) below), which provides a simple M-step. The E-step finds the expectation of the log-likelihood function of θ based on the augmented data Y_{aug}, conditionally on Y_{obs} and $\theta^{(t)}$ from the tth iteration of the EM algorithm. (We follow the convention of using t to index the iteration. This should not be confused with the t-variable or t-model.) Since this log-likelihood is linear in the 'missing' data $(q_1, \ldots, q_n)$, the E-step amounts to calculating

$$w_i^{(t+1)} = E(q_i|y_i, \mu^{(t)}, \Sigma^{(t)}) = \frac{\nu + p}{\nu + d_i^{(t)}}, \qquad i = 1, \ldots, n, \tag{2.2}$$

where $d_i^{(t)} = (y_i - \mu^{(t)})^{\mathrm{T}}(\Sigma^{(t)})^{-1}(y_i - \mu^{(t)})$. Consequently, the standard EM algorithm calculates the $(t+1)$th iterate as

$$\mu^{(t+1)} = \sum_i w_i^{(t+1)} y_i \Big/ \sum_i w_i^{(t+1)}, \tag{2.3}$$

$$\Sigma^{(t+1)} = \frac{1}{n} \sum_i w_i^{(t+1)} (y_i - \mu^{(t+1)})(y_i - \mu^{(t+1)})^{\mathrm{T}}, \tag{2.4}$$

where $w_i^{(t+1)}$ is calculated in equation (2.2). The algorithm then iterates between equations (2.2)–(2.4) until it converges.

Now let us consider a more general data augmentation scheme by multiplying both the numerator and the denominator in equation (2.1) by $|\Sigma|^{-a/2}$, with a being an arbitrary constant, which results in

$$t_p(\mu, \Sigma, \nu) = \mu + \frac{|\Sigma|^{-a/2} \Sigma^{1/2} Z}{\sqrt{q(a)}}, \qquad Z \sim N_p(0, I_p), \quad q(a) \sim |\Sigma|^{-a} \chi_\nu^2 / \nu, \quad Z \perp q(a). \tag{2.5}$$

In other words, we move a portion of the scale factor (this is more transparent for the univariate case, $p = 1$) into the missing data $q(a)$, where the argument a highlights the fact that its distribution now depends on the *working parameter a*. Note that the standard augmentation scheme (2.1) corresponds to $a = 0$ (i.e. $q(0) = q$). Although expression (2.5) is mathematically equivalent to expression (2.1), it provides a different data augmentation scheme because when $q(a)$ is *assumed* to be known it also contributes to the estimation of Σ. In other words, what expression (2.5) accomplishes is to 'transform' part of the M-step into the E-step (or vice versa). For each given a, we can proceed as before to derive the corresponding EM algorithm (which may not be easy to implement) and its rate of convergence as a function of a by treating $Y_{\mathrm{aug}}(a) = \{(y_i, q_i(a)), i = 1, \ldots, n\}$ as the augmented data. In Section 2.4, we shall show that the optimal a that maximizes the speed of the algorithm is $a_{\mathrm{opt}} = 1/(\nu + p)$, a result that is neither obvious nor intuitive (at least to us). Amazingly, the corresponding optimal EM algorithm is not only very easy to implement but also only differs from the standard algorithm (2.2)–(2.4) by a trivial modification, i.e. by replacing the denominator n in equation (2.4) with the sum of the weights (which has already been calculated for equation (2.3)):

$$\Sigma_{\mathrm{opt}}^{(t+1)} = \sum_i w_i^{(t+1)} (y_i - \mu^{(t+1)})(y_i - \mu^{(t+1)})^{\mathrm{T}} \Big/ \sum_i w_i^{(t+1)}. \tag{2.6}$$

This replacement does not change the limit, because $\Sigma_i w_i^{(t+1)} \to n$ as $t \to \infty$. This fact was proved by Kent *et al.* (1994), who used it to construct a modified EM algorithm for fitting t-models and reported that the modified EM algorithm is faster than the standard EM algorithm. It turns out that the optimal EM algorithm given by equations (2.2), (2.3) and (2.6) is identical with their modified EM algorithm, and thus we know not only that their modified EM algorithm has all the properties of the EM algorithm (e.g. monotone convergence in likelihood) but also that it converges faster than the standard algorithm for any t-model being fitted to any data set.

2.3. *Simulation Studies*

To explore the actual gains in computational time of the optimal EM algorithm we conducted several simulations. We first generated 100 observations from each of three distributions:

(a) $N(0, 1)$,
(b) $t_1(0, 1, \nu = 1)$ (i.e. standard Cauchy) and
(c) a mixture of two-thirds $N(0, 1)$ and one-third exponential with mean 3.

We then fitted $t_1(\mu, \Sigma, \nu)$ with $\nu = 1$ and $\nu = 5$ to each data set using both the standard and the optimal EM algorithms. Such simulation configurations are intended to reflect the fact that, in reality, there is no guarantee that the data are from a t-model or even from a symmetric model. We started both algorithms with the same convenient initial values, $\mu^{(0)} = \bar{y}$ and

$$\Sigma^{(0)} = \frac{1}{n}\sum_i (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}$$

(since the variance is not defined for $\nu \leqslant 2$, we use this scalar multiple). We also recorded N_{std} and N_{opt}, the number of iterations required by the standard and optimal algorithms respectively, for achieving $||\theta^{(t)} - \theta^{(t-1)}||^2/||\theta^{(t-1)}||^2 \leqslant 10^{-10}$, where $\theta = (\mu, \Sigma)$. The simulation was repeated 1000 times and the results appear in Fig. 2. (Comparing only the number of iterations is often misleading because different algorithms may take more or less time to complete each iteration; see Table 4 of Section 3.4. In the current case, however, the standard and optimal algorithms clearly require the same amount of computation per iteration.) In all 6000 cases the optimal algorithm was faster than the standard EM algorithm. Generally the improvement was quite significant. In 5997 cases the improvement was greater than 10% and often reached as high as 50% when the Cauchy model ($\nu = 1$) was fitted. Similar empirical evidence is also reported in Arslan *et al.* (1995), using a different convergence criterion. They also provided a quantitative theoretical comparison of the rates of convergence to demonstrate the superiority of the optimal EM algorithm.

A second simulation was run to investigate the improvement in higher dimensions. We fitted a 10-dimensional Cauchy model to 100 observations generated from $t_{10}(0, V, \nu = 1)$, where $V > 0$ was randomly selected at the outset of the simulation as a non-diagonal matrix. Using the same starting values and convergence criterion, N_{std} and N_{opt} were again computed for 1000 data sets. Fig. 3 is a scatterplot of $(N_{\mathrm{std}}, N_{\mathrm{opt}})$ with the improvement $N_{\mathrm{std}}/N_{\mathrm{opt}}$ represented by the broken lines. The improvement of the optimal EM algorithm is dramatic. The standard EM algorithm took at least 6.5 times longer and usually took between 8 and 10 times longer. Comparing this result with the first simulation, we see that the improvement seems to be much more pronounced in higher dimensional problems. We note that when the EM algorithm is slowest and improvement is most useful the gains demonstrated by the optimal algorithm are most striking. It is truly remarkable that such striking gains are obtained without any increase in computation, a true 'free lunch'!

To understand the source of improvement better, Fig. 4 depicts the iterates of the standard and optimal algorithms on the log-likelihood surface for one of the Cauchy data sets generated in the univariate simulation. It is clear that the optimal algorithm

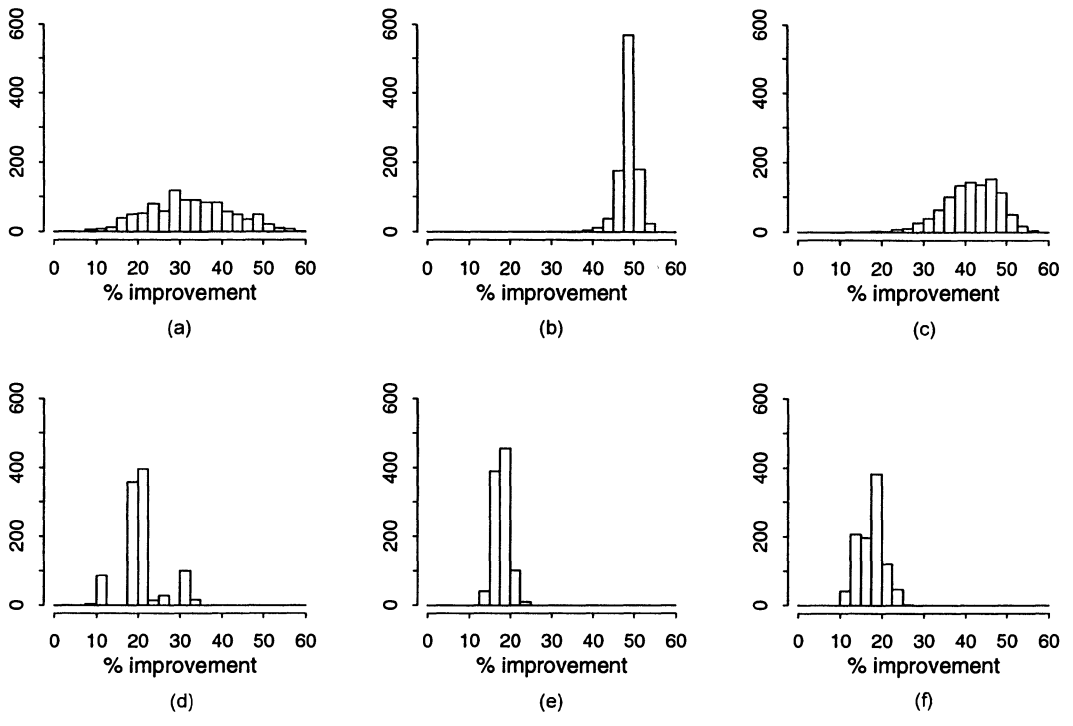

Fig. 2. Percentage improvement of the optimal EM algorithm over the standard EM algorithm for the univariate t-model (each histogram represents 1000 simulated data sets from one of three models to which one of the two t-models was fitted with both the standard and the optimal algorithms; the histograms show the relative improvement in iterations required for convergence (% improvement $= 100(N_{\text{std}} - N_{\text{opt}})/N_{\text{std}}$)): (a) t-model fitted with $\nu = 1$, true data model normal; (b) t-model fitted with $\nu = 1$, true data model Cauchy; (c) t-model fitted with $\nu = 1$, true data model, mixture; (d) t-model fitted with $\nu = 5$, true data model normal; (e) t-model fitted with $\nu = 5$, true data model Cauchy; (f) t-model fitted with $\nu = 5$, true data model, mixture

takes much larger increments in the log-likelihood value, especially during the early iterations. The difference between the two algorithms stems from the expected augmented data log-likelihoods, $Q(\theta|\theta^*) = E[L(\theta|Y_{\text{aug}})|Y_{\text{obs}}, \theta^*]$, which result from the two augmentation schemes. This difference is depicted in Fig. 5 for the same data set used in Fig. 4. From Fig. 5, we see that $Q(\theta|\theta^*)$ resulting from the optimal data augmentation is flatter than $Q(\theta|\theta^*)$ resulting from the standard data augmentation and, hence, approximates $L(\theta|Y_{\text{obs}})$ better. As we show in the following section, this degree of approximation can be ordered using the (observed) Fisher information, which measures the curvature of the log-likelihood surface around the mode.

2.4. *Theoretical Derivations*

In general when constructing an EM algorithm any Y_{aug} is a legitimate data augmentation as long as $Y_{\text{obs}} = \mathcal{M}(Y_{\text{aug}})$ for some (many-to-one) mapping $\mathcal{M}$. Suppose that we have a class of augmentations, $Y_{\text{aug}}(a)$, indexed by a working parameter contained in a set $\mathcal{A}$, such that $Y_{\text{aug}}(a)$ is a legitimate data augmentation for each $a \in \mathcal{A}$. Our goal here is to determine values of a that result in algorithms that

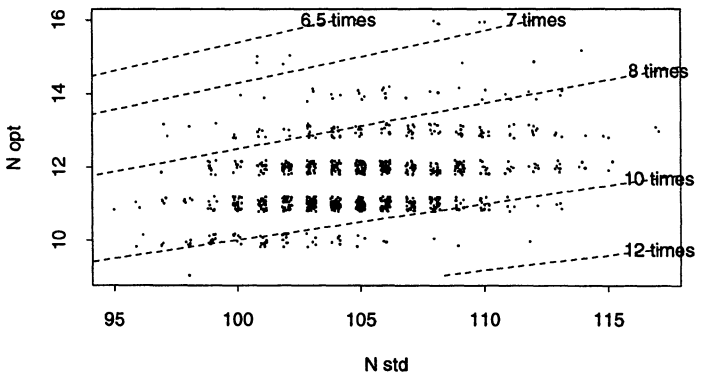

Fig. 3. Improvement of the optimal EM algorithm over the standard algorithm when fitting $t_{10}(\mu, \Sigma, \nu = 1)$: the plot shows the number of iterations required for the standard EM algorithm, N_{std}, and for the optimal EM algorithm, N_{opt}, for each of 1000 simulated 10-variate Cauchy data sets (the points have been jittered using a $U(-0.2, 0.2)$ distribution)

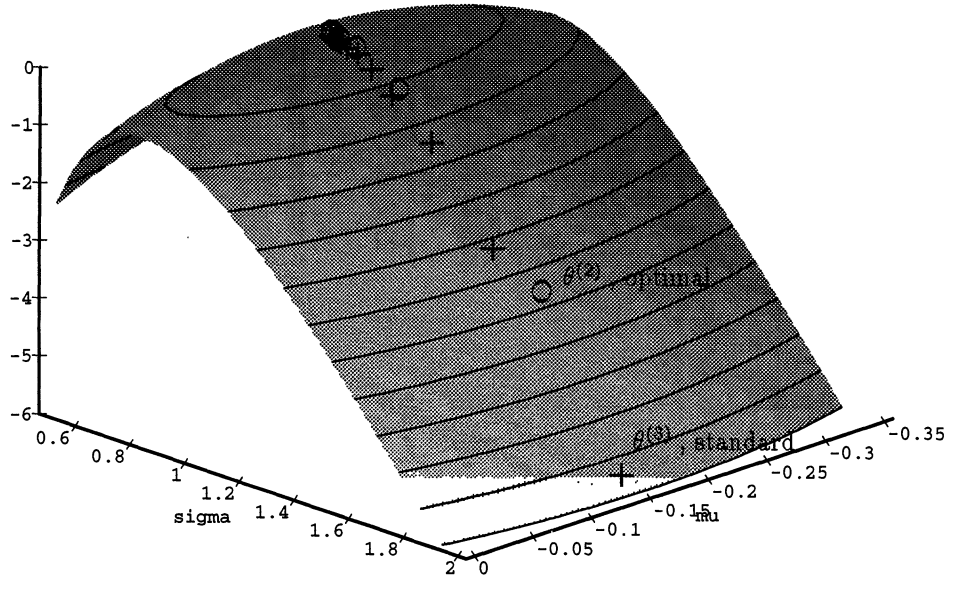

Fig. 4. Comparing the optimal and standard iterative mappings: the figure shows the mappings induced on $L(\theta|Y_{\text{obs}})$ by the standard (+) and the optimal (○) algorithm for a one-dimensional Cauchy data set fitted with $\nu = 1$, starting from the same $\theta^{(0)}$ (not shown); notice how much larger the increments in the log-likelihood values are with the optimal algorithm

are *both quick to converge and easy to implement.* The question of ease of implementation must be considered case by case, so here we confine our attention to the rate of convergence of the EM algorithm as a function of a. From Dempster *et al.* (1977) we know that the matrix rate of the EM algorithm is given by (assuming that the limit of the EM sequence is an interior point)

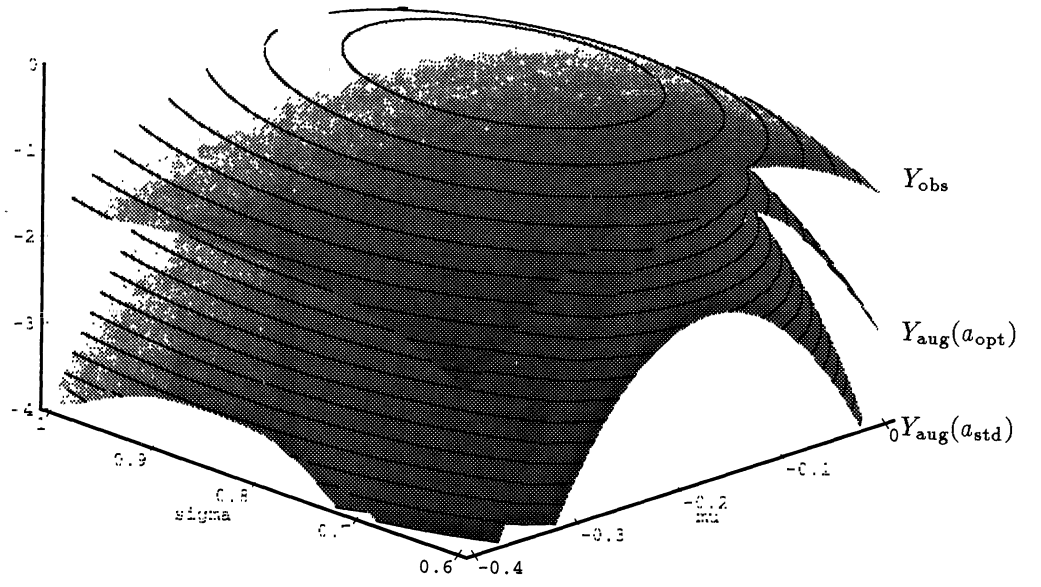

Fig. 5. Comparing the log-likelihoods: the plot shows $L(\theta|Y_{obs})$, as well as $E[L(\theta|Y_{aug})|Y_{obs}, \theta^*]$ for both the standard and the optimal augmentations (each adjusted by their maximum value for comparison); notice that the optimal augmentation results in a flatter log-likelihood that better approximates $L(\theta|Y_{obs})$

$$\mathrm{DM}^{\mathrm{EM}}(a) = I - I_{obs}\, I_{aug}^{-1}(a),$$

where I is the identity matrix,

$$I_{aug}(a) = E\left[-\frac{\partial^2[\log f\{Y_{aug}(a)|\theta\}]}{\partial\theta\partial\theta^{\mathrm{T}}}\,\middle|\, Y_{obs}, \theta\right]\Bigg|_{\theta=\theta^*} \tag{2.7}$$

is the expected augmented information and

$$I_{obs} = -\frac{\partial^2 L(\theta|Y_{obs})}{\partial\theta\partial\theta^{\mathrm{T}}}\Bigg|_{\theta=\theta^*} \tag{2.8}$$

is the observed information matrix, which is positive semidefinite when θ^* is a (local) MLE. The largest eigenvalue of $\mathrm{DM}^{\mathrm{EM}}(a)$, denoted by $r(a)$, is known as the (global) rate of convergence of the EM algorithm (e.g. Meng and Rubin (1994)). Since large values of $r(a)$ result in *slower* algorithms, Meng (1994) defined $s(a) = 1 - r(a)$ as the global speed of the algorithm. Note that $s(a)$ is the smallest eigenvalue of the speed matrix $S^{\mathrm{EM}}(a) = I_{obs} I_{aug}^{-1}(a)$.

Therefore, our goal here is to make $s(a)$ as large as possible as a function of a. Since I_{obs} is independent of the data augmentation scheme, it is enough to search for a small $I_{aug}(a)$ in the sense of a positive semidefinite ordering, as proved in the following theorem. The same ordering was used by Fessler and Hero (1994) for the SAGE algorithm.

Theorem 1. If $I_{\text{aug}}(a) \geqslant I_{\text{aug}}(a')$, i.e. $I_{\text{aug}}(a) - I_{\text{aug}}(a')$ is positive semidefinite, then $s(a) \leqslant s(a')$.

Proof. If $I_{\text{obs}} > 0$, then $s(a)$ is the smallest eigenvalue of $B(a) \equiv I_{\text{obs}}^{1/2} I_{\text{aug}}^{-1}(a) I_{\text{obs}}^{1/2}$. But $I_{\text{aug}}(a) \geqslant I_{\text{aug}}(a')$ implies $B(a) \leqslant B(a')$ (e.g. Horn and Johnson (1985), p. 470), and thus the result follows trivially from the Courant–Fischer representation: $s(a) = \min_{b^{\mathrm{T}} b=1}\{b^{\mathrm{T}} B(a) b\}$. If $|I_{\text{obs}}| = 0$, then $s(a) = s(a') = 0$. □

We now apply theorem 1 to show that replacing equation (2.4) with equation (2.6) results in the optimal algorithm in the class defined by equation (2.5). For a fixed a, the log-likelihood for (μ, Σ) based on $Y_{\text{aug}}(a)$ is

$$L(\mu, \Sigma | Y_{\text{aug}}(a)) = \frac{n}{2}\{a(p+\nu) - 1\} \log |\Sigma| - \frac{1}{2}|\Sigma|^a \sum_i q_i(a)\{\nu + (\bar{y}_w - \mu)^{\mathrm{T}} \Sigma^{-1}(\bar{y}_w - \mu) + \operatorname{tr}(\Sigma^{-1} S_w)\}, \tag{2.9}$$

where

$$\bar{y}_w = \sum_i q_i(a) y_i \Big/ \sum_i q_i(a), \qquad S_w = \sum_i q_i(a)(y_i - \bar{y}_w)(y_i - \bar{y}_w)^{\mathrm{T}} \Big/ \sum_i q_i(a). \tag{2.10}$$

It follows immediately that the MLE given $Y_{\text{aug}}(a)$ for μ is $\bar{y}_w$. To obtain the MLE of Σ, we differentiate equation (2.9) with respect to the elements of $\Psi = \Sigma^{-1}$, which yields the following score equation (a detailed derivation is given in van Dyk (1995)):

$$\frac{\partial}{\partial \Psi} L(\mu, \Sigma | Y_{\text{aug}}(a)) = -\frac{n}{2}\{a(p+\nu) - 1\}\, \mathcal{D}(\Sigma) - \frac{n}{2}|\Sigma|^a\, \bar{q}(a)\{(\bar{y}_w - \mu)(\bar{y}_w - \mu)^{\mathrm{T}} + \mathcal{D}(S_w)\} + \frac{na}{2}\bar{q}(a)|\Sigma|^a\{\nu + (\bar{y}_w - \mu)^{\mathrm{T}} \Sigma^{-1}(\bar{y}_w - \mu) + \operatorname{tr}(\Sigma^{-1} S_w)\}\, \mathcal{D}(\Sigma) = 0, \tag{2.11}$$

where $\bar{q}(a) = \sum_{i=1}^n q_i(a)/n$ and $\mathcal{D}(A) = 2A - \operatorname{diag}(A)$, i.e. $\mathcal{D}(A)$ doubles all the elements of A except the diagonal ones. Evaluating equation (2.11) at the MLE of $\mu = \bar{y}_w$ and replacing $\mathcal{D}(\Sigma)$ with Σ and $\mathcal{D}(S_w)$ with S_w, since we may find the root of equation (2.11) element by element, we see that the MLE of Σ satisfies

$$\frac{a(p+\nu) - 1}{|\Sigma|^a\, \bar{q}(a)}\, \Sigma + S_w = a\{\nu + \operatorname{tr}(\Sigma^{-1} S_w)\}\Sigma. \tag{2.12}$$

Solving equation (2.12) with arbitrary a is quite difficult if not impossible, but there are two values of a that make it trivial to solve. One is $a = 0$, corresponding to the standard augmentation scheme, which yields $\Sigma = \bar{q}(0) S_w$ and thus equation (2.4). The other is $a_{\text{opt}} = 1/(\nu + p)$, with which the first term on the left-hand side of equation (2.12) is 0, and thus the solution Σ must be proportional to S_w. It is then easy to verify that the proportionality constant must be 1, and therefore the MLE of Σ is S_w, which leads to the corresponding M-step in equation (2.6). It is arguably a miracle that equation (2.12) can be solved analytically for the optimal a, but for

(almost) no other a. It is also remarkable that the optimal a does not depend on the particular data set. We believe that these findings are not coincidental, although we are still searching for a simple explanation.

To show that a_{opt} indeed yields the best possible rate of convergence under the augmentation scheme (2.5), we apply theorem 1 and need to verify only that $I_{\text{aug}}(a) \geqslant I_{\text{aug}}(a_{\text{opt}})$ for all a. Using techniques similar to those used to derive equation (2.11), one can verify that

$$I_{\text{aug}}(a) - I_{\text{aug}}(a_{\text{opt}}) = (\nu + p)\left(a - \frac{1}{\nu + p}\right)^2 C_n \qquad \text{for any } a,$$

where

$$C_n = \frac{n}{2}\begin{pmatrix} 0 & 0 \\ 0 & \varsigma\varsigma^{\mathrm{T}} \end{pmatrix},$$

$\varsigma = \text{vec}\{\mathcal{D}(\Sigma^*)\}$, $\text{vec}(A)$ stacks a $d \times d$ symmetric matrix A into a $d(d+1)/2$-dimensional vector row by row without duplicating the off-diagonal elements and Σ^* is the MLE of Σ. It follows immediately that $a_{\text{opt}} = 1/(\nu + p)$ minimizes the augmented information since C_n does not depend on a. With the note that the E-step for the optimal EM algorithm only differs from the standard E-step of equation (2.2) by a scale factor that is independent of i and, thus, is irrelevant for equations (2.3) and (2.6), this completes our proof that replacing equation (2.4) by equation (2.6) results in a uniformly faster EM algorithm regardless of ν, p or Y_{obs}.

In this derivation, we could order $I_{\text{aug}}(a)$ with the positive semidefinite ordering, in which case it defines an ordering of the data augmentation schemes. When $I_{\text{aug}}(a) \geqslant I_{\text{aug}}(a')$, we may say that the augmentation $Y_{\text{aug}}(a')$ is *nested* in $Y_{\text{aug}}(a)$. In such cases, when we write $I - \text{DM}^{\text{EM}}(a) \equiv S^{\text{EM}}(a)$ as

$$S^{\text{EM}}(a) = \{I_{\text{obs}}\, I_{\text{aug}}^{-1}(a')\}\{I_{\text{aug}}(a')\, I_{\text{aug}}^{-1}(a)\} \equiv S^{\text{EM}}(a')\, R(a', a), \tag{2.13}$$

we can view $R(a', a)$ as the speed of an EM algorithm with 'observed data' $Y_{\text{aug}}(a')$ and augmented data $Y_{\text{aug}}(a)$. Of course, two augmentations need not be nested (i.e. $I_{\text{aug}}(a) - I_{\text{aug}}(a')$ may be neither positive nor negative semidefinite). In such cases $R(a', a) \equiv I_{\text{aug}}(a')\, I_{\text{aug}}^{-1}(a)$ can be viewed as the relative augmented information. Intuitively, if $R(a', a)$ is 'small', $Y_{\text{aug}}(a')$ will result in a faster algorithm than $Y_{\text{aug}}(a)$ will, and, if it is large, the opposite will be true. When the augmentations are not nested, theorem 1 does not apply but $R(a', a)$ is still useful for searching for efficient algorithms; see Meng and van Dyk (1997) for a detailed illustration with the mixed effects model.

3. ALTERNATING EXPECTATION–CONDITIONAL MAXIMIZATION ALGORITHM

3.1. *Data Augmentation and Model Reduction*

Data augmentation is the key idea underlying the EM algorithm, and in this section we shall show its power in constructing efficient algorithms when coupled with model reduction. We recall that the EM algorithm augments the observed data

Y_{obs} to the larger Y_{aug}. Starting with an initial value $\theta^{(0)} \in \Theta$, it then finds θ^*, a stationary point of $L(\theta|Y_{obs})$, by iterating the following two steps ($t = 0, 1, \ldots$):

E-step—impute the augmented data log-likelihood $L(\theta|Y_{aug})$ by

$$Q(\theta|\theta^{(t)}) = E[L(\theta|Y_{aug})|Y_{obs}, \theta^{(t)}]; \tag{3.1}$$

M-step—determine $\theta^{(t+1)}$ by maximizing the imputed log-likelihood $Q(\theta|\theta^{(t)})$,

$$Q(\theta^{(t+1)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)}), \qquad \text{for all } \theta \in \Theta. \tag{3.2}$$

The idea is to select Y_{aug} so that $\theta^{(t+1)}$ is easy to compute, thereby providing a simple, stable, though sometimes slow, algorithm.

Although the notion of data augmentation is appealing to statisticians, it is sometimes mysterious to others who are puzzled by the notion of using 'data' that one does not have. For some, a more familiar approach for fitting a complicated model is to break it into several smaller models, i.e. instead of augmenting the data to 'match' the model we can reduce the model to 'match' the data. For example, we may be able to find a partition $\theta = (\vartheta_1, \vartheta_2)$ such that $L(\theta|Y_{obs})$ is easy to maximize for ϑ_1 given ϑ_2 and vice versa. In such cases, we can iterate between these two conditional (or constrained) maximizations until convergence to maximize $L(\theta|Y_{obs})$. This model reduction method is well known in the optimization literature as the cyclic coordinate ascent method (e.g. Zangwill (1969)) and can also be viewed as the Gauss–Seidel method applied in an appropriate order to the score equations. It turns out that simple partitions of θ are not enough for certain statistical models (e.g. log-linear models for contingency tables) to achieve simple conditional maximizations (CMs), and thus Meng and Rubin (1993) studied more general model reduction strategies by introducing a set of $S \geqslant 1$ (vector) constraint functions $G = \{g_s(\theta), s = 1, \ldots, S\}$ that are 'space filling', a condition that ensures that the entire parameter space be searched after the completion of the S CMs defined by G. Specifically, at each iteration of this CM-algorithm, we carry out the following S CM-steps in turn for $s = 1, \ldots, S$:

sth CM-step—determine $\theta^{(t+s/S)}$ by maximizing the constrained observed data log-likelihood

$$L(\theta^{(t+s/S)}|Y_{obs}) \geqslant L(\theta|Y_{obs}), \qquad \text{for all } \theta \in \Theta_s^{(t)} \equiv \{\theta \in \Theta: g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})\}. \tag{3.3}$$

The next iterate is then defined as $\theta^{(t+1)} = \theta^{(t+S/S)}$.

The main purpose of Meng and Rubin (1993) was to demonstrate the benefit of combining data augmentation and model reduction by introducing the expectation–conditional maximization (ECM) algorithm, which replaces the original M-step of the EM algorithm by a set of CM-steps analogous to step (3.3):

sth CM-step—determine $\theta^{(t+s/S)}$ by maximizing the constrained imputed log-likelihood

$$Q(\theta^{(t+s/S)}|\theta^{(t)}) \geqslant Q(\theta|\theta^{(t)}), \qquad \text{for all } \theta \in \Theta_s^{(t)} \equiv \{\theta \in \Theta: g_s(\theta) = g_s(\theta^{(t+(s-1)/S)})\}. \tag{3.4}$$

The rationale for this replacement is that in some applications of the EM algorithm the M-step requires iteration but with an appropriately chosen G all the CM-steps defined by expression (3.4) can have analytic solutions or are at least much simpler to implement than the original M-step. Meng and Rubin (1993) gave three examples to illustrate this point and also proved that the ECM algorithm retains the stable convergence properties of the EM algorithm, in particular the monotone convergence of the likelihood values. In addition, they defined the multicycle ECM (MCECM) algorithm, which inserts an E-step before some of the CM-steps. The hope is that by updating the imputed log-likelihood more often one can achieve faster convergence (surprisingly, although the MCECM algorithm always involves more computation per iteration than the corresponding ECM algorithm, the former does not necessarily have a faster rate of convergence; see Meng (1994) for a counter-example).

A trade-off of the ECM algorithm, when compared with the EM algorithm, is that it usually converges more slowly in terms of the global speed of convergence defined in Section 2.4 (surprisingly, again, this is not always the case; see Meng (1994)). We emphasize that this does not necessarily imply that the ECM algorithm takes longer to converge in real time since the actual time taken for running the M-step and the CM-steps is not taken into account in the calculation of the speed of convergence and can be quite different, especially if the M-step requires iteration. An even more relevant comparison for individual users should also include the human investment for implementing the algorithms; after all, it is this comparison that makes the EM algorithm more popular for statisticians in some settings than, say, Newton–Raphson iteration. Having said this, however, it is obviously desirable to speed up an algorithm if we can with little increase in human effort; Section 2 showed that this is possible even for a popular algorithm. Two recent generalizations of the ECM algorithm, the ECME algorithm (Liu and Rubin, 1994) and the SAGE algorithm (Fessler and Hero, 1994) also represent such efforts.

Specifically, Liu and Rubin (1994) recognized that in some applications of the ECM algorithm the implementation of some of the CM-steps requires similar computations for maximizing the observed data log-likelihood and for maximizing the expected augmented data log-likelihood under the same constraint (i.e. similar computation for step (3.3) and for step (3.4)). In such cases, one is tempted to perform step (3.3) rather than step (3.4), as the former operates on the actual log-likelihood (i.e. no augmentation) and thus should lead to faster convergence. Liu and Rubin (1994, 1995) provided numerical examples to demonstrate the gain in central processor unit (CPU) time, sometimes by a factor of 10, which resulted from this modification. They thus defined the ECME algorithm whose CM-steps maximize *either* the imputed log-likelihood (i.e. $Q(\theta|\theta^{(t)})$) *or* the actual log-likelihood (i.e. $L(\theta|Y_{\text{obs}})$).

Fessler and Hero (1994) developed SAGE without knowledge of the ECM or ECME algorithms and motivated their algorithm from a different perspective from that of Meng and Rubin (1993). As we have discussed, the ECM algorithm starts with EM and then incorporates model reduction (i.e. CM) into the M-step to simplify the computation. In contrast, SAGE starts with the CM algorithm (i.e. break the problem into several smaller problems by conditioning sequentially on a subset of the parameters) and then uses data augmentation (i.e. EM) to simplify the computation of each reduced problem, which may still have no analytic solution. Because each of the reduced problems considers the likelihood as a function of a

different subset of the parameters, it is natural to use a different data augmentation scheme for each of the corresponding EM algorithms. In some settings, this attempt to simplify computations also turns out to be very useful for speeding up the algorithm, because the amount of data augmentation needed for each of the smaller problems can be less than that needed for the original big problem.

Our purpose here is to provide a general framework that combines the ECM and SAGE algorithms. The AECM algorithm, defined in Section 3.2, allows the data augmentation scheme to vary within and between iterations, a key ingredient of SAGE, and allows the model reduction strategy to go beyond a simple partition of the parameters, a key feature of the ECM algorithm. After presenting theoretical results on the AECM algorithm in Section 3.3, we introduce in Section 3.4 an AECM algorithm for fitting the t-model with unknown degrees of freedom. In Section 3.5, we illustrate the use of working parameters together with the AECM algorithm for image reconstruction with the Poisson model.

3.2. *Alternating Expectation–Conditional Maximization Algorithm*

To present our algorithm in its most general form, we need to extend and standardize the indexing system that has been commonly used in the EM literature. Specifically, we need to develop the notion of a 'cycle' in between a 'step' and an 'iteration'.

Definition 1. A *cycle* consists of an E-step followed by an ordered set of CM-steps, the last of which will be followed immediately by a new E-step (which is itself the beginning of the next cycle). An *iteration* consists of one or more cycles.

For example, in the ECM algorithm, an iteration is the same as a cycle, but for the MCECM algorithm an iteration consists of multiple cycles (hence its name). In what follows, we use t, c and s to index iteration, cycle and step respectively, as illustrated in the following scheme:

$$\underbrace{\underbrace{\text{E-step, } \{s\text{th CM-step, } s = 1, \ldots, S^{[c_t+1]}\}}_{(c_t+1)\text{th cycle}}, \ldots, \underbrace{\text{E-step, } \{s\text{th CM-step, } s = 1, \ldots, S^{[c_{t+1}]}\}}_{(c_{t+1})\text{th cycle}}}_{t\text{th iteration}}$$

The flexibility of our general algorithm comes from allowing the data augmentation scheme, as well as the set of constraint functions, to depend on the cycle index. At the $(c+1)$th cycle, we first choose the augmented data $Y_{\text{aug}}^{[c+1]}$ and the set of constraining functions $G^{[c+1]} = \{g_s^{[c+1]}(\theta),\ s = 1, \ldots, S^{[c+1]}\}$. The $(c+1)$th cycle of the AECM algorithm then consists of

E-step—compute $Q^{[c+1]}(\theta|\theta^{[c]}) = E[L(\theta|Y_{\text{aug}}^{[c+1]})|Y_{\text{obs}}, \theta^{[c]}]$

and $S^{[c+1]}$ CM-steps

sth CM-step ($s = 1, \ldots, S^{[c+1]}$)—calculate $\theta^{[c+s/S^{[c+1]}]}$ such that

$$Q^{[c+1]}(\theta^{[c+s/S^{[c+1]}]}|\theta^{[c]}) \geqslant Q^{[c+1]}(\theta|\theta^{[c]})$$

$$\text{for all } \theta \in \Theta_s^{[c+1]} \equiv \{\theta \in \Theta: g_s^{[c+1]}(\theta) = g_s^{[c+1]}(\theta^{[c+(s-1)/S^{[c+1]}]})\}. \tag{3.5}$$

The output of the $(c+1)$th cycle is then defined as $\theta^{[c+S^{[c+1]}/S^{[c+1]}]} = \theta^{[c+1]}$.

It is clear that without restrictions on $G^{[c]} = \{g_s^{[c]}, s = 1, \ldots, S^{[c]}\}$ there is no reason to hope that the AECM algorithm will converge properly. The condition needed here is the *space filling* condition which Meng and Rubin (1993) used for the ECM algorithm. Intuitively, this condition requires that, after a set of constrained maximizations, we will have searched in all directions (radiating from a particular origin) of the parameter space. Mathematically, the space filling condition holds for $G^{[c]} = \{g_s^{[c]}, s = 1, \ldots, S^{[c]}\}$ at a particular $\theta' \in \Theta$ if and only if (see Meng and Rubin (1993))

$$\bigcap_{s=1}^{S^{[c]}} J_s^{[c]}(\theta') = \{0\}, \qquad \text{where } J_s^{[c]}(\theta') = \{\nabla g_s^{[c]}(\theta')\lambda: \lambda \in \mathbb{R}^{d_s^{[c]}}\} \tag{3.6}$$

is the column space of the gradient (denoted by ∇) of the $d_s^{[c]}$-dimensional vector function $g_s^{[c]}(\theta')$, $\nabla g_s^{[c]}(\theta')$. (We always assume that $g_s^{[c]}(\theta)$ is differentiable and $\nabla g_s^{[c]}(\theta)$ is of full rank at each $\theta \in \Theta_0$, the interior of Θ, to avoid unnecessary technical complications.) For the ECM algorithm, each iteration consists of one cycle, and $G^{[c]}$ does not depend on the cycle index. Thus, it was sufficient in Meng and Rubin (1993) only to consider constraint (3.6) for one cycle. With the AECM algorithm, however, it is possible that the space filling condition will not hold for every cycle, i.e. it may be useful (for the efficiency of the algorithm) to allow the space filling condition to be satisfied only after several cycles (e.g. as in Fessler and Hero (1994)). More precisely, in contrast with constraint (3.6), we may now only require

$$\bigcap_{c=c_1}^{c_2} \bigcap_{s=1}^{S^{[c]}} J_s^{[c]}(\theta') = \{0\}, \tag{3.7}$$

where c_1 and c_2 are two generic indices. To facilitate proper monitoring of convergence (see Meng and van Dyk (1995)), we define an iteration of the AECM algorithm as the smallest set of consecutive cycles such that constraint (3.7) holds, instead of defining cycles as iterations. More precisely, using parentheses to index iteration and square brackets to index cycle, we define an AECM iteration sequence $\{\theta^{(t)}, t \geqslant 0\}$ as a subsequence of the sequence generated by the output of each cycle $\{\theta^{[c]}, c \geqslant 0\}$ such that $\theta^{(t+1)} = \theta^{[c_{t+1}]}$ if

$$\bigcap_{c=c_t+1}^{c_{t+1}} \bigcap_{s=1}^{S^{[c]}} J_s^{[c]}(\theta^{[c_t]}) = \{0\}, \quad \text{but} \bigcap_{c=c_t+1}^{c_{t+1}-1} \bigcap_{s=1}^{S^{[c]}} J_s^{[c]}(\theta^{[c_t]}) \neq \{0\}. \tag{3.8}$$

In other words, we consider a set of consecutive cycles to form an iteration of the AECM algorithm if and only if the last cycle of the set has completed the search of the parameter space, starting from the previous iteration (not cycle) output, in the sense of completing the space filling requirement. Since the cycles are time ordered, there is a unique sequence $\{c_t; t \geqslant 0\}$ that defines the iteration sequence; obviously we

also need to assume that this unique sequence consists of infinitely many elements. All these requirements become transparent when we consider the conditions needed for the convergence of the Gibbs sampler, which can be viewed as the stochastic counterpart of the AECM algorithm as we shall discuss in Section 4, for which we need to require that the resulting Markov chain has positive probability to visit every state infinitely many times.

Table 2 provides an overview of how the AECM algorithm generalizes several existing algorithms including the CM, EM, ECM, MCECM, SAGE and ECME algorithms, where the definition of ECME is modified from Liu and Rubin's (1994) to guarantee the monotonicity of the likelihood along any ECME sequence (see the remark below).

3.3. *Convergence Results for Alternating Expectation–Conditional Maximization Algorithm*

We now show that the sequence $\{\theta^{(t)}\}$ from an AECM algorithm increases $L(\theta|Y_{\text{obs}})$ at each iteration and that under standard regularity conditions the sequence converges to a stationary point of $L(\theta|Y_{\text{obs}})$. These results are the counterparts of the results for the EM algorithm (Dempster *et al.*, 1977; Wu, 1983) and for the ECM algorithm (Meng and Rubin, 1993), and provide more complete results for ECME and SAGE.

Our first and most important result states that the AECM algorithm, like all its predecessors, maintains monotonic convergence of the likelihood values. Note that this result does not require the space filling condition.

Theorem 2. Any AECM sequence increases (or maintains) $L(\theta|Y_{\text{obs}})$ at every cycle and thus increases (or maintains) $L(\theta|Y_{\text{obs}})$ at every iteration.

Proof. By the construction of the AECM algorithm, we have

$$Q^{[c+1]}(\theta^{[c+s/S^{[c+1]}]}|\theta^{[c]}) \geq Q^{[c+1]}(\theta^{[c+(s-1)/S^{[c+1]}]}|\theta^{[c]}), \qquad \text{for } s = 1, \ldots, S^{[c+1]}$$

and thus

$$Q^{[c+1]}(\theta^{[c+1]}|\theta^{[c]}) \geq Q^{[c+1]}(\theta^{[c]}|\theta^{[c]}), \tag{3.9}$$

TABLE 2
Special cases of the AECM algorithm†

Case of AECM	$Y_{\text{aug}}^{[c]}$	$C^{(t)}$	$S^{[c]}$	$G^{[c]}$
CM	Y_{obs}	1	S	G
EM	Y_{aug}	1	1	None
ECM	Y_{aug}	1	S	G
MCECM	Y_{aug}	C	1	$G_{(c \text{ modula } C)}$
ECME	$\begin{cases} Y_{\text{aug}}, c \text{ odd} \\ Y_{\text{obs}}, c \text{ even} \end{cases}$	2	$\begin{cases} S_1, c \text{ odd} \\ S_2, c \text{ even} \end{cases}$	$\begin{cases} G_1, c \text{ odd} \\ G_2, c \text{ even} \end{cases}$
SAGE	$Y_{\text{aug}}^{[c]}$	$C^{(t)}$	1	Partition of θ

†The table records the data augmentation scheme ($Y_{\text{aug}}^{[c]}$), the number of cycles per iteration ($C^{(t)}$), the number of CM-steps per cycle ($S^{[c]}$) and the constraint functions ($G^{[c]}$). When the index c (or t) is suppressed, the quantity is fixed between cycles (or iterations).

which implies that $L(\theta^{[c+1]}|Y_{\text{obs}}) \geqslant L(\theta^{[c]}|Y_{\text{obs}})$, using the same argument as in the proof of Dempster *et al.* (1977) for the EM algorithm. □

We remark here that the statement

$$Q(\theta_1|\tilde{\theta}) \geqslant Q(\theta_2|\tilde{\theta}) \Longrightarrow L(\theta_1|Y_{\text{obs}}) \geqslant L(\theta_2|Y_{\text{obs}}) \tag{3.10}$$

is true in general only when $\theta_2 = \tilde{\theta}$ (e.g. as in inequality (3.9)). In Liu and Rubin's (1994) proof of their theorem 1, expression (3.10) was implicitly assumed when $\theta_2 \neq \tilde{\theta}$, which invalidates the proof. This problem is easily avoided if we always perform an E-step whenever we change the data augmentation scheme. Our modification to the ECME algorithm which always performs those CM-steps with no augmentation after those with augmentation accomplishes this because when there is no augmentation the E-step is an identity mapping.

To prove that $\{\theta^{(t)}, t \geqslant 0\}$ converges to

$$\theta^* \in \mathcal{L} = \left\{\theta \in \Theta_0 \colon \frac{\partial L(\theta|Y_{\text{obs}})}{\partial \theta} = 0\right\}$$

where Θ_0 is the interior of Θ (i.e. θ^* is an interior stationary point), several standard regularity conditions used by Wu (1983) and by Meng and Rubin (1993) for the EM and ECM algorithms respectively are required; in particular, we assume Wu's (1983) conditions (6)–(10). The proof of the following theorem relies on the global convergence theorem (see Wu (1983)) and the space filling condition, and it is similar to the proof of theorem 2 of Meng and Rubin (1993); for brevity, details are omitted here but can be found in Meng and van Dyk (1995).

Theorem 3. In addition to Wu's (1983) regularity conditions (6)–(10), suppose that

(a) all the CMs in step (3.5) are unique and
(b) the AECM iteration mapping, $M_{(t)}^{\text{AECM}}\colon \theta^{(t)} \to \theta^{(t+1)}$, does not depend on t.

Then all the limit points of an AECM sequence $\{\theta^{(t)}, t \geqslant 0\}$ are stationary points of $L(\theta|Y_{\text{obs}})$.

As discussed in Meng and Rubin (1993), the uniqueness condition (a) is often satisfied in practice (e.g. when the CMs are in closed form), but even this condition can be eliminated if we force $\theta^{[c+s/S^{[c+1]}]} = \theta^{[c+(s-1)/S^{[c+1]}]}$ whenever there is no increase in $Q^{[c+1]}(\theta|\theta^{[c]})$ at the sth CM-step within the $(c+1)$th cycle. Other conditions are also possible to ensure the result as discussed in Meng and Rubin (1993) and Lange (1995a). Corollary 1 of Meng and Rubin (1993) also holds here, i.e. if $L(\theta|Y_{\text{obs}})$ is unimodal with θ^* being the only stationary point then, under the conditions of theorem 3, any AECM sequence will converge to θ^* starting from any $\theta^{(0)} \in \Theta_0$.

Finally, if in addition to assuming that $M_{(t)}^{\text{AECM}}$ does not depend on t we assume that each iteration has the same set of cycles, say $\{c_1, \ldots, c_C\}$, we have the following result regarding the rate of convergence of the AECM algorithm, the proof of which is essentially the same as Meng's (1994) proof of the rate of convergence for a multi-cycle ECM algorithm, and thus is omitted here. Note that this result applies only when θ^* is an interior point of Θ, as is the case for the corresponding EM and ECM rate results (see Dempster *et al.* (1977) and Meng (1994)).

Theorem 4. Suppose that the AECM iteration mapping is a composition of C fixed cycle mappings, all the CMs in step (3.5) satisfy the Lagrange multiplier equations and $\theta^{[c+s/S^{[c+1]}]} \to \theta^*$ as $c \to \infty$. Then the matrix rate of convergence of the AECM iteration is

$$\mathrm{DM}^{\mathrm{AECM}} = \prod_{c=1}^{C} \left\{ I - I_{\mathrm{obs}}(I_{\mathrm{aug}}^{[c]})^{-1} \left(I - \prod_{s=1}^{S^{[c]}} P_s^{[c]} \right) \right\}, \tag{3.11}$$

where the products are ordered from left to right with increasing index, I_{obs} is given in equation (2.8),

$$I_{\mathrm{aug}}^{[c]} = E\left[-\frac{\partial^2 \{\log f(Y_{\mathrm{aug}}^{[c]} | \theta)\}}{\partial\theta\partial\theta^{\mathrm{T}}} \,\middle|\, Y_{\mathrm{obs}}, \theta \right]\Bigg|_{\theta=\theta^*},$$

$$P_s^{[c]} = \nabla_s^{[c]} \{ (\nabla_s^{[c]})^{\mathrm{T}} (I_{\mathrm{aug}}^{[c]})^{-1} \nabla_s^{[c]} \}^{-1} (\nabla_s^{[c]})^{\mathrm{T}} (I_{\mathrm{aug}}^{[c]})^{-1},$$

$\nabla_s^{[c]} = \nabla g_s^{[c]}(\theta^*)$, $G^{[c]} = \{g_s^{[c]}(\theta),\ s = 1, \ldots, S^{[c]}\}$ and $Y_{\mathrm{aug}}^{[c]}$ is determined by the cth cycle, $c = 1, \ldots, C$.

This theorem provides a unified expression for the results on the rate of convergence of all special cases listed in Table 2. When $C = 1$, the supplemented ECM algorithm (van Dyk *et al.*, 1995), which combines the supplemented EM algorithm (Meng and Rubin, 1991) and the ECM algorithm, uses equation (3.11) to calculate the asymptotic variance–covariance matrix of θ^*, I_{obs}^{-1}, as a function of $\mathrm{DM}^{\mathrm{ECM}}$, $\Pi_{s=1}^{S} P_s$ and I_{aug}. When $C > 1$, there is often a corresponding algorithm with $C = 1$ that can be used in conjunction with a supplemented ECM algorithm to calculate I_{obs}^{-1}, once θ^* has been obtained (see van Dyk (1995) for more discussion).

3.4. *Example: Fitting t-models with Unknown Degrees of Freedom*

To construct fast algorithms for fitting a t-model with unknown degrees of freedom ν, we first reduce the model by breaking the parameter space into two parts to use a two-cycle algorithm. In the first cycle, we update (μ, Σ) given ν and, in the second cycle, we update ν given (μ, Σ). In other words, we choose $\Theta^{[c]} = \{\theta \equiv (\mu, \Sigma, \nu) \in \Theta\colon \nu = \nu^{[c-1]}\}$ for c odd and $\Theta^{[c]} = \{\theta \in \Theta\colon (\mu, \Sigma) = (\mu^{[c-1]}, \Sigma^{[c-1]})\}$ for c even. The resulting AECM algorithm has two cycles in each iteration and one CM-step in each cycle. Given ν, as was discussed in Section 2.2, there are two choices for the data augmentation which result in CM-steps that are simple to compute. The first is the standard augmentation, $\{(y_i, q_i),\ i = 1, \ldots, n\}$, and the second is the optimal augmentation, $\{(y_i, q_i(a_{\mathrm{opt}}^{[c]})),\ i = 1, \ldots, n\}$, where (y_i, q_i) is as in model (2.1), $q_i(a) = |\Sigma|^{-a} q_i$ and $a_{\mathrm{opt}}^{[c]} = 1/(\nu^{[c]} + p)$. Since $\nu^{[c]}$ changes with c the optimal data augmentation scheme is not fixed but rather is a function of the cycle and the iteration (and thus it does not fit the ECME framework). There are also two useful data augmentation schemes when we condition on μ and Σ (i.e. the even-numbered cycles): the standard augmentation, $\{(y_i, q_i),\ i = 1, \ldots, n\}$, and no augmentation, $\{y_i,\ i = 1, \ldots, n\}$; the latter was used by Liu and Rubin (1994) in their ECME implementation.

In conjunction with the standard data augmentation in the odd cycles (which update (μ, Σ)), the two augmentation schemes in the even cycles result in the MCECM

and ECM algorithms and were compared by Liu and Rubin (1994, 1995). Here we compare the MCECM and ECME algorithms with two new algorithms (AECM 1 and AECM 2 respectively) which result from replacing the standard augmentation with the optimal augmentation when updating (μ, Σ) (see Table 3). Implementation of all four algorithms is straightforward. The odd-numbered cycles are conditional on the current iterate $\nu^{[c]}$ and are implemented exactly as described in Section 2.3 with ν replaced by $\nu^{[c]}$. With the change in notation from $(t+1)$ to $[c+1]$, the E-step is given in equation (2.2), and for the standard augmentation the CM-step is given in equations (2.3) and (2.4). For the optimal augmentation we replace equation (2.4) with equation (2.6). Regardless of the data augmentation, the even cycles require numerical optimization of $Q^{[c+2]}(\theta|\theta^{[c+1]})$, where c is even. Specifically, when using $Y_{\text{aug}} = \{(y_i, q_i), i = 1, \ldots, n\}$, we first execute an E-step which sets $w_i^{[c+2]} = (\nu^{[c]} + p)/(\nu^{[c]} + d_i^{[c+1]})$ for each i, and then set $\nu^{[c+2]}$ equal to the solution of the equation

$$-\phi\left(\frac{\nu}{2}\right) + \log\left(\frac{\nu}{2}\right) + \phi\left(\frac{\nu^{[c]}+p}{2}\right) - \log\left(\frac{\nu^{[c]}+p}{2}\right) + \frac{1}{n}\sum_{i=1}^{n}(\log w_i^{[c+2]} - w_i^{[c+2]}) + 1 = 0, \tag{3.12}$$

where $\phi(\cdot)$ is the digamma function. Likewise, when using $Y_{\text{aug}} = \{y_i, i = 1, \ldots, n\}$, ν is updated by setting $\nu^{[c+2]}$ equal to the solution of the equation

$$-\phi\left(\frac{\nu}{2}\right) + \log\left(\frac{\nu}{2}\right) + \phi\left(\frac{\nu+p}{2}\right) - \log\left(\frac{\nu+p}{2}\right) + \frac{1}{n}\sum_{i=1}^{n}\{\log \tilde{w}_i^{[c+2]}(\nu) - \tilde{w}_i^{[c+2]}(\nu)\} + 1 = 0, \tag{3.13}$$

where $\tilde{w}_i^{[c+2]}(\nu) = (\nu + p)/(\nu + d_i^{[c+1]})$. Equations (3.12) and (3.13) are special cases of equations (27) and (30) given in Liu and Rubin (1995).

To compare the performance of the four algorithms described in Table 3, we fitted a 10-dimensional t-model to 100 observations generated from $t_{10}(0, V, \nu = 1)$, where $V > 0$ was selected at the outset of the simulation as a non-diagonal matrix. The half-interval method (e.g. Carnahan *et al.* (1969)) was used to solve equations (3.12) and (3.13). Using the same convergence criterion and starting values for (μ, Σ) as

TABLE 3

Four AECM algorithms used to fit the multivariate t-distribution with unknown degrees of freedom

Algorithm	*Model reduction (CM-steps):* *Update μ and Σ*		*Update ν*	
	Data augmentation (E-step): $\{(y_i, q_i)\}$	$\{(y_i, q_i(a_{\text{opt}}^{(i)})\}$	$\{(y_i, q_i)\}$	$\{(y_i)\}$
MCECM	X		X	
ECME	X			X
AECM 1		X	X	
AECM 2		X		X

described in the simulation in Section 2.3 and the starting value $\nu^{(0)} = 10$, the number of iterations required for convergence was recorded for each of the four algorithms. Fig. 6 contains scatterplots which compare AECM 1 with each of the other three algorithms. As we see, AECM 1 was 8–12 times faster than either the MCECM or ECME algorithm. Note that the cost per iteration is less for AECM 1 and for the MCECM algorithm than for the ECME algorithm, mainly because solving equation (3.13) requires more computation than solving equation (3.12). Moreover, AECM 2 (the combination of the optimal EM algorithm and ECME) was only slightly more efficient than AECM 1 in terms of the number of iterations required, and less efficient in terms of actual computer time.

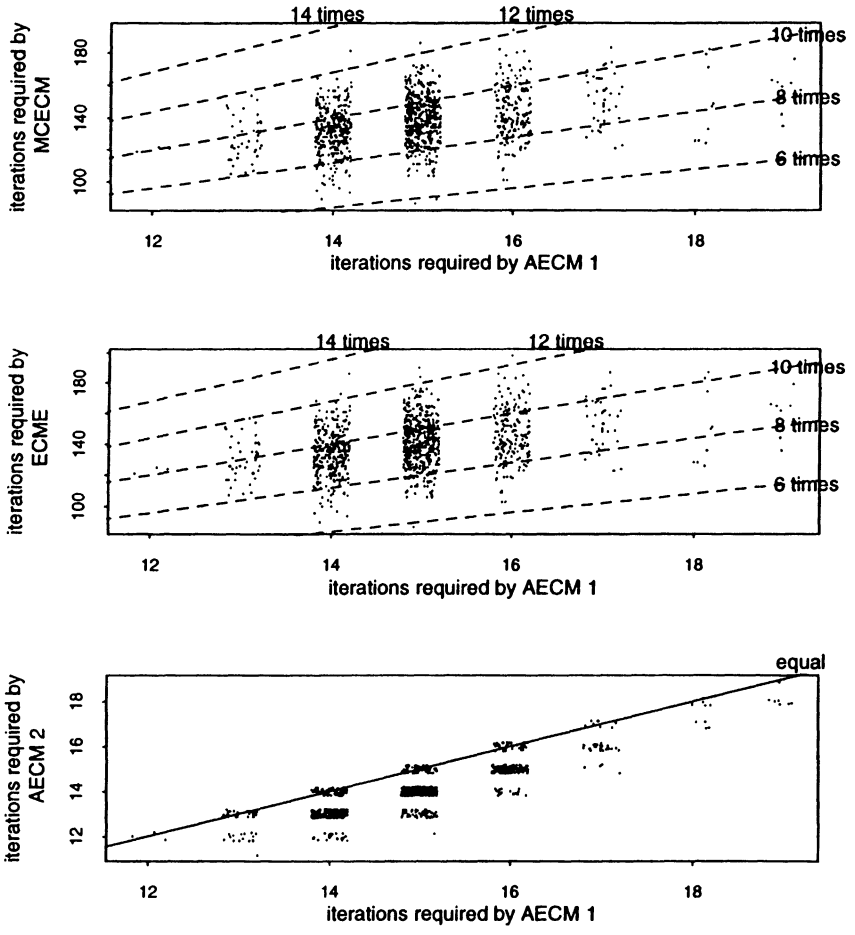

Fig. 6. Comparing AECM algorithms for fitting the multivariate t-distribution with unknown degrees of freedom: the scatterplots compare the number of iterations required for convergence by each of the MCECM, ECME and AECM 2 algorithms with AECM 1; AECM 1 and AECM 2 use the optimal data augmentation of Section 2.3 for (μ, Σ) and perform very well; ECME and AECM 2 use Y_{obs} in place of Y_{aug} when updating ν and show only a small improvement over MCECM and AECM 1 respectively; this improvement is washed out by the increased computation time required per iteration by AECM 2 (the points were jittered as in Fig. 2)

To investigate the relative merit of the four algorithms in Table 3 in the presence of missing values among the y_is, we replicated some of the studies in Liu and Rubin (1995). The ECME algorithm was run with the code provided by Liu and the other three algorithms were run with minor modifications of his ECME code. The second and third columns of Table 4 show the performance of the four algorithms for fitting a four-dimensional data set (from Shih and Weisberg (1986)) with only a few missing values. For this problem, ECME was the slowest and AECM 1 is the fastest in terms of CPU time, differing by a factor of more than 10. The second data set is a two-dimensional artificial data set used by Murray in his discussion of Dempster *et al.* (1977) to illustrate the problem with multimodal likelihood functions. This data set which has many missing values among the y_i results in a three-mode t-likelihood (with $\mu^* = (0, 0)$)—two modes are global ($L = -45.5543$) and one is local ($L = -45.7713$), as reported in Liu and Rubin (1995). Contrary to Table 3 of Liu and Rubin (1995), ECME requires the most CPU time in this example as well. (The computer that we used is much faster than that used to produce Liu and Rubin's (1995) Table 3.) Thus, for fitting a t-model with unknown degrees of freedom, we recommend AECM 1 unless there are many missing values among the y_i in which case AECM 2 may be somewhat more efficient than AECM 1.

3.5. *Example: Image Reconstruction under Poisson Model*

The EM algorithm has been applied to image reconstruction problems since the work of Shepp and Vardi (1982) (see also Vardi *et al.* (1985) and Lange and Carson (1984)), but slow convergence has hampered its usefulness in this application. Various attempts have been made to speed up this EM implementation, as reviewed in Fessler and Hero (1994), who themselves have written a series of papers on the subject. We view Fessler and Hero's work to be statistically most promising, as they clearly emphasize the importance of speeding up the algorithm without sacrificing its simplicity or stability, and skilfully use data augmentation techniques to construct efficient algorithms. Indeed, after we submitted the first version of this paper in August 1995, we learned from Jeffery Fessler (in his comments on our manuscript) that our following generalization of Fessler and Hero (1994) had been considered in Fessler and Hero (1995) and that our anticipation of its relative superiority in

TABLE 4
Comparison of the performance of the algorithms described in Table 3 on the two data sets in Liu and Rubin (1995)†

Algorithm	*Results for Shih and Weisberg's data*		*Results for Murray's data*		
	CPU time (s)	*Iterations*	*CPU time (s)*	*Iterations*	$L(\theta^*\|Y_{obs})$
MCECM	12	48	38	266	−45.5543
ECME	95	40	77	61	−45.7713
AECM 1	8	31	34	236	−45.5543
AECM 2	50	21	22	19	−45.7713

†The algorithms were run on a Sun SPARC station 1.

practice has been confirmed by empirical evaluations. We thus have eliminated our original discussions on the practical issues of implementing this algorithm and refer interested readers to Fessler and Hero (1995) for these details, as well as for modifications of the algorithm that incorporate smoothing penalty functions (i.e. prior distributions) to improve the generally poor quality of the MLE reconstruction (also see Silverman *et al.* (1990)). Our purpose here is to illustrate how to construct such an algorithm by coupling the working parameter idea of Section 2 with alternating data augmentation schemes. We also compare it theoretically with several other algorithms, including Newton–Raphson iteration, in terms of step sizes.

To illustrate, we follow Fessler and Hero's (1994) notation. They considered data generated from

$$Y_n \sim \text{Poisson}\left(\sum_{k=1}^{p} a_{nk}\lambda_k + r_n\right), \qquad \text{independent for } n = 1, \ldots, N, \tag{3.14}$$

where k $(= 1, \ldots, p)$ indexes the pixels, n $(= 1, \ldots, N)$ indexes the bins (e.g. a pair of detectors), a_{nk} is the conditional probability that an emission from pixel k will be detected by bin n, r_n is the background rate and $\{a_{nk}, r_n\}$ are known. We want to estimate the emission densities $\lambda = (\lambda_1, \ldots, \lambda_p)^T$ by maximum likelihood. The computational burden for the MLE stems from the large sizes of N and p, both of which are often in the tens of thousands. The standard EM implementation for this problem is to use

$$Y_{\text{aug}}^{\text{EM}} = \{Y_{nk}, R_n, 1 \leqslant k \leqslant p, 1 \leqslant n \leqslant N\}, \tag{3.15}$$

where $Y_{nk} \sim \text{Poisson}(a_{nk}\lambda_k)$, $R_n \sim \text{Poisson}(r_n)$ and all the Poisson variables are independent. In other words, for each bin the augmented data consist of the number of emissions originating from each pixel as well as from the background. Clearly, equation (3.15) is a massive augmentation over the observed data (3.14) and thus leads to a slow EM algorithm.

To improve on this, Fessler and Hero (1994) first broke the problem into p small problems (one can also group pixels into small groups). Given $\lambda_j, j \neq k$, expression (3.14) becomes

$$Y_n \sim \text{Poisson}\{a_{nk}\lambda_k + r_n + \Lambda_{nk}(\lambda)\}, \qquad \text{independent for } n = 1, \ldots, N, \tag{3.16}$$

where $\Lambda_{nk}(\lambda) = \Sigma_{j \neq k} a_{nj}\lambda_j$ is known. Since there is only one unknown parameter, λ_k, in expression (3.16), it is possible to find a data augmentation scheme that is much smaller in terms of the Fisher information for λ_k than equation (3.15) is but sufficiently large for easy EM-type implementation. Specifically, we can use

$$Y_{\text{aug}}^{[k]}(\beta_k) = \{Z_{nk}(\beta_k), B_{nk}(\beta_k), 1 \leqslant n \leqslant N\}, \tag{3.17}$$

where

$$Z_{nk}(\beta_k) \sim \text{Poisson}\{a_{nk}(\lambda_k + \beta_k)\}, \tag{3.18}$$

$$B_{nk}(\beta_k) \sim \text{Poisson}\{r_n - a_{nk}\beta_k + \Lambda_{nk}(\lambda)\}, \tag{3.19}$$

all Poisson variables are independent and β_k is a non-negative *working parameter*

such that the Poisson parameter in expression (3.19) is non-negative. The rationale behind this augmentation scheme is simply that we can decompose $Y_n = Z_{nk}(\beta_k) + B_{nk}(\beta_k)$, and if $Z_{nk}(\beta_k)$ were observed then the MLE for λ_k would be

$$\hat{\lambda}_k = \max\left(\frac{Z_{\cdot k}}{a_{\cdot k}} - \beta_k, 0\right), \tag{3.20}$$

where $Z_{\cdot k} = \Sigma_{n=1}^{N} Z_{nk}(\beta_k)$ and $a_{\cdot k} = \Sigma_{n=1}^{N} a_{nk}$.

This defines the M-step of the EM algorithm for the reduced model (3.16), or equivalently the CM-step of the ECM algorithm for updating λ_k given λ_j, $j \neq k$, under the original model (3.14). Since $Z_{\cdot k}$ is unobserved, and it is the augmented data sufficient statistic for λ, the E-step replaces $Z_{\cdot k}$ in equation (3.20) by its conditional expectation given $Y_{\text{obs}} = \{y_n, 1 \leqslant n \leqslant N\}$ and the most up-to-date estimate of λ, say, $\lambda^{[c]}$:

$$Z_{\cdot k}^{[c+1]} \equiv E(Z_{\cdot k} | Y_{\text{obs}}, \lambda_k^{[c]}) = (\lambda_k^{[c]} + \beta_k) \sum_{n=1}^{N} \frac{a_{nk} y_n}{a_{nk}\lambda_k^{[c]} + r_n + \Lambda_{nk}(\lambda^{[c]})}. \tag{3.21}$$

Once we have completed an E-step and an M-step for λ_k, which corresponds to the kth cycle of the AECM algorithm, we move on to the next cycle for λ_{k+1}, and so forth until we complete an iteration of AECM when all of $\lambda = (\lambda_1, \ldots, \lambda_p)^{\mathrm{T}}$ is updated. Since each AECM iteration consists of p cycles, we have $\lambda^{(t)} = \lambda^{[pt]}$, using the notation of Section 3.2. Since each λ_k, $k = 1, \ldots, p$, is only updated at the kth cycle of each iteration, we have

$$\lambda^{[tp+k-1]} = (\lambda_1^{(t+1)}, \ldots, \lambda_{k-1}^{(t+1)}, \lambda_k^{(t)}, \ldots, \lambda_p^{(t)})^{\mathrm{T}}, \tag{3.22}$$

which is the most up-to-date estimate of λ when $\lambda_k^{(t)}$ is being updated to $\lambda_k^{(t+1)}$. For this reason, c will be set to $tp + k - 1$ in the following derivation.

The unsettled piece in the above construction is the choice of $\{\beta_k, 1 \leqslant k \leqslant p\}$. Fessler and Hero (1994) provided the above construction with $\beta_k = z_k$, where

$$z_k = \min_{n:\, a_{nk} > 0} (r_n / a_{nk}), \qquad k = 1, \ldots, p. \tag{3.23}$$

Although this value of β_k produces a faster algorithm than the ECM implementation without introducing the working parameter (which is the same as setting $\beta_k = 0$ in this problem; this ECM implementation differs from the standard EM implementation only in terms of how the previous iterate of λ is used in equation (3.21)), it can still be improved on, perhaps substantially. In particular, it is easy to verify that under the *constrained model* (3.16) the augmented data Fisher information based on equation (3.17) is

$$I_{\text{aug}}^{[k]}(\beta_k) = \frac{a_{\cdot k}}{\lambda_k^* + \beta_k}, \tag{3.24}$$

assuming the MLE for λ_k, $\lambda_k^* > 0$. On the basis of equation (3.24), Fessler and Hero (1994) concluded that it is much better to use $\beta_k = z_k$ than $\beta_k = 0$, because it reduces the augmented data Fisher information. However, from equation (3.24) it is clear that $\beta_k = z_k$ is not the value of the working parameter that will minimize the augmented data information under the constraint (see expression (3.19))

$$\beta_k \leqslant \frac{r_n + \Lambda_{nk}(\lambda)}{a_{nk}}, \qquad \text{for all } n\text{: } a_{nk} > 0. \tag{3.25}$$

The value that minimizes the augmented data information under this constraint is

$$\beta_k^{\text{OPT}} = \min_{n:a_{nk}>0} \left\{ \frac{r_n + \Lambda_{nk}(\lambda)}{a_{nk}} \right\}. \tag{3.26}$$

Fessler and Hero's (1994) choice of z_k relies on the fact that the total background rates correspond to roughly 10% of the total count in a typical positron emission tomography study and resulted in a significant improvement in the convergence rate of the algorithm. Therefore, using β_k^{OPT} instead of z_k can produce an even more substantial improvement, because

$$r_n + \Lambda_{nk}(\lambda) = r_n + \sum_{j \neq k} a_{nj}\lambda_j$$

is the expected total emission received by detector n except those originating at pixel k, a quantity that should be much larger than r_n. Fessler and Hero (1995) call the contribution from all other pixels 'pseudobackground' events.

To see more clearly the advantage of using $\beta_k = \beta_k^{\text{OPT}}$ over using $\beta_k = z_k$ (and $\beta_k = 0$), we provide the following theoretical comparisons, which may be more appealing to those who are more comfortable with numerical analysis than with statistics (e.g. comparing the Fisher information). Specifically, let

$$\bar{y}_n(\lambda) = \sum_{k=1}^{p} a_{nk}\lambda_k + r_n, \qquad n = 1, \ldots, N;$$

we have

$$\beta_k^{\text{OPT}} = \min_{n:a_{nk}>0} \{\bar{y}_n(\lambda)/a_{nk}\} - \lambda_k.$$

Thus, the iteration mapping resulting from using $\beta_k = \beta_k^{\text{OPT}}$, as determined from equations (3.20) and (3.21), is given by (recall that $c = tp + k - 1$)

$$\lambda_k^{(t+1)} = \max\left[\lambda_k^{(t)} - \{a_{\cdot k} - e_k(\lambda^{[c]})\}\frac{f_k(\lambda^{[c]})}{a_{\cdot k}}, 0\right], \tag{3.27}$$

where

$$e_k(\lambda) = \sum_{n=1}^{N} \frac{a_{nk}y_n}{\bar{y}_n(\lambda)}$$

and

$$f_k(\lambda) = \min_{n:a_{nk}>0} \{\bar{y}_n(\lambda)/a_{nk}\} \qquad k = 1, \ldots, p.$$

It is easy to verify from Fessler and Hero (1994) that the Newton–Raphson algorithm, the EM algorithm with $\beta_k = 0$, and Fessler and Hero's algorithm with $\beta_k = z_k$

(each applied in *conjunction* with the model reduction scheme given by expression (3.16)) all have the following form of mapping:

$$\lambda_k^{(t+1)} = \max[\lambda_k^{(t)} - \{a_{.k} - e_k(\lambda^{[c]})\}\, S_k(\lambda^{[c]}), 0], \tag{3.28}$$

where, for the Newton–Raphson, EM and Fessler and Hero's (1994) original algorithms,

$$\left.\begin{aligned} S_k^{\mathrm{NR}}(\lambda) &= \left\{\sum_{n=1}^{N} a_{nk}^2 y_n / \bar{y}_n^2(\lambda)\right\}^{-1}, \\ S_k^{\mathrm{EM}}(\lambda) &= \lambda_k / a_{.k}, \\ S_k^{\mathrm{FH}}(\lambda) &= (\lambda_k + z_k)/a_{.k}. \end{aligned}\right\} \tag{3.29}$$

From equation (3.27) for the optimal algorithm,

$$S_k^{\mathrm{OPT}}(\lambda) = \frac{f_k(\lambda)}{a_{.k}} = \frac{1}{a_{.k}} \min_{n:a_{nk}>0} \left\{ \frac{\Lambda_{nk}(\lambda)}{a_{nk}} + \frac{r_n}{a_{nk}} + \lambda_k \right\} \geqslant \frac{1}{a_{.k}} \min_{n:a_{nk}>0} \left\{ \frac{\Lambda_{nk}(\lambda)}{a_{nk}} \right\} + \frac{z_k + \lambda_k}{a_{.k}}, \tag{3.30}$$

and thus, excluding the mathematical possibility of equality,

$$S_k^{\mathrm{OPT}}(\lambda) > S_k^{\mathrm{FH}}(\lambda) > S_k^{\mathrm{EM}}(\lambda), \tag{3.31}$$

so each improvement increases the step size and thus achieves faster convergence. (Since $S_k^{\mathrm{EM}}(\lambda) \propto \lambda_k$, the EM implementation is slow because the step size will be very small for the $\lambda_k^{(t)}$s that converge to 0 or nearly to 0.) Furthermore, on convergence, $a_{.k} \geqslant e_k(\lambda^*)$, and thus

$$S_k^{\mathrm{NR}}(\lambda^*) \geqslant \left[\max_{n:a_{nk}>0} \left\{ \frac{a_{nk}}{\bar{y}_n(\lambda^*)} \right\} \sum_{n=1}^{N} \frac{a_{nk} y_n}{\bar{y}_n(\lambda^*)} \right]^{-1} = \frac{f_k(\lambda^*)}{e_k(\lambda^*)} \geqslant \frac{f_k(\lambda^*)}{a_{.k}} = S_k^{\mathrm{OPT}}(\lambda^*). \tag{3.32}$$

In other words, the optimal algorithm is the closest to the Newton–Raphson algorithm near convergence (EM-type algorithms cannot be made identical with Newton–Raphson algorithms, so the inequality in expression (3.32) is generally strict), but with more stable behaviour than the Newton–Raphson algorithm in the early iterations because of the guaranteed monotone convergence in likelihood. The key idea here, from an algorithmic point of view, is to search for algorithms, within the class of algorithms with monotone convergence in likelihood, that have as large a step size as possible. The AECM formulation provides a more flexible setting for this search than did its predecessors.

To emphasize the simplicity of the resulting algorithm, we provide the following program outline, modified from Fessler and Hero (1995).

$$\textit{Initialize: } \bar{y}_n := \sum_{k=1}^{p} a_{nk}\lambda_k^{(0)} + r_n, \quad n = 1, \ldots, N, \quad a_{.k} := \sum_{n=1}^{N} a_{nk}, \quad k = 1, \ldots, p,$$

$$\textit{for } t = 0, 1, \ldots, \{$$

$$\textit{for } k = 1, \ldots, p \{ f_k := \min_{n:a_{nk}>0} \left(\frac{\bar{y}_n}{a_{nk}} \right); \quad e_k := \sum_{n=1}^{N} \frac{a_{nk} y_n}{\bar{y}_n},$$

$$\lambda_k^{(t+1)} := \max \left\{ \lambda_k^{(t)} - \left(1 - \frac{e_k}{a_{.k}} \right) f_k, 0 \right\},$$

$$\bar{y}_n := \bar{y}_n + (\lambda_k^{(t+1)} - \lambda_k^{(t)}) a_{nk} \quad \textit{for each } n \textit{ such that } a_{nk} > 0$$

$$\}$$

$$\}.$$

Although this algorithm offers substantial gains over both the EM and the original SAGE algorithm of Fessler and Hero (1994), there are probably other AECM implementations that have better theoretical rates of convergence. This algorithm was derived by optimizing the EM algorithm for each of the *constrained models* given in expression (3.16) separately. In theory we should choose β_k, $k = 1, \ldots, p$, to minimize the spectral radius of $\mathrm{DM}^{\mathrm{AECM}}$ as given in equation (3.11). Such an optimization should also take into account the order in which the cycles are implemented. Of course, there is no guarantee that such improvements in the asymptotic rate of convergence will reduce the required computation time (e.g. van Dyk and Meng (1997)), especially if λ^* is on the boundary of Θ. Nor is there any guarantee that these algorithms will be as easy to implement as the current algorithm.

4. EPILOGUE: SINGING EM ALGORITHM WITH RANDOM NOTES

Wherever we have presented the working parameter approach, we have always been asked the same question, 'would the same idea work for the Gibbs sampler?'. Indeed, the problem of monitoring convergence of a Gibbs sampler, or more generally Markov chain Monte Carlo algorithms, is of central importance in the routine use of these powerful tools. Speeding up these stochastic algorithms not only means saving time but also more importantly increases the number of properly converged samplers in practice, where an algorithm is often stopped simply because a user cannot afford to run it longer.

It has long been observed that there is an intrinsic link between EM-type algorithms and stochastic algorithms like the Gibbs sampler. Tanner and Wong's (1987) data augmentation algorithm, for example, was directly motivated by the EM algorithm. Inspired by the observation that Tanner and Wong's algorithm is a Gibbs sampler with two conditional distributions, Meng (1990) proposed the partitioned ECM algorithm, which is a special case of the ECM algorithm in which the constraint functions form a simple partition of the parameter space, just as one does with the standard Gibbs sampler. The subsequent formulation of the ECM algorithm with more general constraint functions in turn suggested a similar generalization of the standard Gibbs sampler, which led to Bayesian iterative proportional fitting (IPF) for posterior analysis of contingency tables (see Gelman *et al.* (1995), section

14.7). Liu and Rubin (1994) drew a parallel relationship between the ECME algorithm and the collapsed or block Gibbs sampler (e.g. Liu *et al.* (1994) and Liu (1994)). It is thus natural to expect a more general formulation of the Gibbs sampler that adopts the alternating data augmentation scheme underlying the SAGE and AECM algorithms. Table 5 summarizes the parallel evolution, in the order of generality, of these two powerful families of statistical algorithms.

Given this intrinsic link, we expect that the working parameter approach will also be useful to accelerate stochastic algorithms. Interestingly, in the context of the Gibbs sampler, the working parameter approach becomes less mysterious because there is only one operator (i.e. random draws), in contrast with the separation of the 'expectation operator' and the 'maximization operator' within EM-type algorithms. For example, for the t-model, using the augmented data $\{(y_i, q_i(a) = |\Sigma|^{-a} q_i), i = 1, \ldots, n\}$ for simulating the posterior distribution of the parameter is equivalent to performing a transformation of the random variable before implementing the Gibbs sampler, because both Σ and q are random variables. Our key message here is that such a transformation can depend on a working parameter that is introduced solely for the search for an efficient algorithm, i.e. although the working parameter is not identifiable (in fact, is not visible) from the marginal distribution of interest, it indexes a class of joint distributions to which we can apply the Gibbs sampler. In fact, the working parameter can be a 'working functional'. For example, we could consider $\{(y_i, \alpha(\mu, \Sigma)q_i), i = 1, \ldots, n\}$ as the augmented data for the t-model and search in some class for an α that results in an efficient algorithm (indeed, α could even be random). We have not explored these directions but warn against too complex a search as this can lead to many 'optimal' results with little practical significance. We also think that directly searching for efficient Gibbs sampler implementations is more difficult than for efficient EM-type algorithms, since the former requires us to deal with a rate of convergence involving a whole distribution, rather than just the information matrices evaluated at the point of convergence. Thus, it might be useful to search for efficient EM-type algorithms first, even if we are not interested in calculating posterior modes (which are useful for constructing starting values for stochastic simulation; see Gelman and Rubin (1992)). The insight gained from this search can then be used to suggest similar data augmentation schemes for efficient Gibbs samplers since, as Smith and Roberts (1993) emphasized, whenever we can implement the EM algorithm, we can implement a corresponding Gibbs sampler by replacing the E- and M-steps with Monte Carlo draws.

We hope that we have provided some evidence that there is indeed an 'algorithmic Utopia': algorithms that are *simple, stable and fast*, and efficient data augmentation coupled with model reduction is a promising route to reach this Utopia.

TABLE 5
EM-type algorithms and their stochastic counterparts

Deterministic algorithm	*Stochastic algorithm*
EM	Tanner and Wong's algorithm
Partitioned ECM	Standard Gibbs sampler
ECM	Generalized Gibbs sampler (e.g. Bayesian IPF)
ECME	Collapsed or block Gibbs sampler
SAGE or AECM	More flexible Gibbs sampler

ACKNOWLEDGEMENTS

The authorship is alphabetical. The research was supported in part by National Science Foundation (NSF) grants DMS 92-04504, DMS 95-05043 and DMS 96-26691, and in part by the US Bureau of the Census through a contract with the National Opinion Research Center at the University of Chicago. It was also supported in part by National Security Agency grant MDA 904-96-1-0007. The manuscript was prepared using computer facilities supported in part by several NSF grants awarded to the Department of Statistics at the University of Chicago, by the University of Chicago Block Fund and by a MacArthur Fellowship granted to van Dyk by Kalamazoo College. We thank J. Kent, D. Tyler and Y. Vardi, and J. Fessler and A. Hero for their inspiring papers and for helpful communications including copies of papers. We also thank Y. Amit, H. Chernoff, N. Cressie, A. Gelman, A. Kong, R. Little, D. Pollard, D. Rubin, H. Stern, B. Sutradhar, W. Wong and A. Zaslavsky for helpful exchanges and C. Liu for comments as well as the use of his ECME computer code. Finally, we thank the Royal Statistical Society Research Section and their reviewers for the quick handling of our submission, for helpful comments and instructions on presentation and most importantly for providing us the opportunity of presenting this work.

REFERENCES

Arslan, O., Constable, P. D. L. and Kent, J. T. (1995) Convergence behaviour of the EM algorithm for the multivariate *t*-distribution. *Communs Statist. Theory Meth.*, **24**, 2981–3000.

Baum, L. E., Petrie, T., Soules, G. and Weiss, N. (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, **41**, 164–171.

Carnahan, B., Luther, H. A. and Wilkes, J. O. (1969) *Applied Numerical Methods*. New York: Wiley.

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

van Dyk, D. A. (1995) Construction, implementation, and theory of algorithms based on data augmentation and model reduction. *PhD Thesis*. Department of Statistics, University of Chicago, Chicago.

van Dyk, D. A. and Meng, X. L. (1997) On the orderings and groupings of conditional maximizations within ECM-type algorithms. *J. Comput. Graph. Statist.*, to be published.

van Dyk, D. A., Meng, X. L. and Rubin, D. B. (1995) Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statist. Sin.*, **5**, 55–75.

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

——(1995) Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithm. *IEEE Trans. Image Process.*, **4**, 1417–1438.

Fisher, R. A. (1925) Theory of statistical estimation. *Proc. Camb. Phil. Soc.*, **22**, 700–725.

Gelman, A., Carlin, J., Stern, H. and Rubin, D. B. (1995) *Bayesian Data Analysis*. London: Chapman and Hall.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–472.

Hartley, H. O. (1958) Maximum likelihood procedures for incomplete data. *Biometrics*, **14**, 174–194.

Horn, R. A. and Johnson, C. R. (1985) *Matrix Analysis*. New York: Cambridge University Press.

Jamshidian, M. and Jennrich, R. I. (1993) Conjugate gradient acceleration of the EM algorithm. *J. Am. Statist. Ass.*, **88**, 221–228.

Kent, J. T., Tyler, D. E. and Vardi, Y. (1994) A curious likelihood identity for the multivariate *t*-distribution. *Communs Statist. Simuln Computn*, **23**, 441–453.

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

——(1995b) A quasi-Newtonian acceleration of the EM algorithm. *Statist. Sin.*, **5**, 1–18.

Lange, K. and Carson, R. (1984) EM reconstruction algorithms for emission and transmission tomography. *J. Comput. Assist. Tomogr.*, **8**, 306–316.

Lange, K., Little, R. J. A. and Taylor, J. M. G. (1989) Robust statistical modeling using the *t*-distribution. *J. Am. Statist. Ass.*, **84**, 881–896.

Lansky, D. and Casella, G. (1990) Improving the EM algorithm. In *Computing Science and Statistics: Proc. Symp. Interface*, pp. 420–424. New York: Springer.

Liu, C. and Rubin, D. B. (1994) The ECME algorithm: a simple extension of EM and ECM with fast monotone convergence. *Biometrika*, **81**, 633–648.

——(1995) ML estimation of the *t*-distribution using EM and its extensions, ECM and ECME. *Statist. Sin.*, **5**, 19–40.

Liu, J. S. (1994) The collapsed Gibbs sampler in Bayesian computations with application to a gene regulation problem. *J. Am. Statist. Ass.*, **89**, 958–966.

Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

McKendrik, A. G. (1926) Applications of mathematics to medical problems. *Proc. Edinb. Math. Soc.*, **44**, 98–130.

Meng, X. L. (1990) Towards complete results for some incomplete-data problems. *PhD Thesis*. Department of Statistics, Harvard University, Cambridge.

——(1994) On the rate of convergence of the ECM algorithm. *Ann. Statist.*, **22**, 326–339.

Meng, X. L. and van Dyk, D. A. (1995) The EM algorithm—an old folk song sung to a fast new tune. *Technical Report 408*. Department of Statistics, University of Chicago, Chicago.

——(1997) Fast EM implementations for random effects models. Submitted to *J. R. Statist. Soc.* B.

Meng, X. L. and Pedlow, S. (1992) EM: a bibliographic review with missing articles. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 24–27.

Meng, X. L. and Rubin, D. B. (1991) Using EM to obtain asymptotic variance-covariance matrices: the SEM algorithm. *J. Am. Statist. Ass.*, **86**, 899–909.

——(1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika*, **80**, 267–278.

——(1994) On the global and componentwise rates of convergence of the EM algorithm. *Lin. Alg. Applic.*, **199**, 413–425.

Shepp, L. A. and Vardi, Y. (1982) Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Image Process.*, **2**, 113–122.

Shih, W. J. and Weisberg, S. (1986) Assessing influence in multiple linear regression with incomplete data. *Technometrics*, **28**, 231–239.

Silverman, B. W., Jones, M. C., Wilson, J. D. and Nychka, D. W. (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). *J. R. Statist. Soc.* B, **52**, 271–324.

Smith, A. F. M. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 3–23.

Stigler, S. (1994) Citation patterns in the journals of statistics and probability. *Statist. Sci.*, **9**, 94–108.

Sundberg, R. (1972) Maximum likelihood theory and applications for distributions generated when observing a function of an exponential variable. *PhD Thesis*. Institute of Mathematics and Statistics, Stockholm University, Stockholm.

——(1974) Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, **1**, 49–58.

——(1976) An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communs Statist. Simuln Computn*, **5**, 55–64.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Statist. Ass.*, **82**, 528–550.

Vardi, Y. and Lee, D. (1993) From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (with discussion). *J. R. Statist. Soc.* B, **55**, 569–612.

Vardi, Y., Shepp, L. A. and Kaufman, L. (1985) A statistical model for positron emission tomography. *J. Am. Statist. Ass.*, **80**, 8–19.

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

Zangwill, W. (1969) *Nonlinear Programming—a Unified Approach*. Englewood Cliffs: Prentice Hall.

DISCUSSION OF THE PAPER BY MENG AND VAN DYK

The following contributors were invited to lead the discussion.

Donald B. Rubin (Harvard University, Cambridge): I congratulate Professor Meng and Professor van Dyk on a 'scoring' contribution exquisitely timed to celebrate the 20th anniversary of the EM paper (Dempster *et al.*, 1977). I am particularly honoured to be invited to initiate this discussion because Meng is my former student and van Dyk my 'grand student' and colleague at Harvard, and I consequently feel the pride that a father feels for superb accomplishments of his progeny.

My comments on this contribution have a past component, dealing with work before Dempster *et al.* (1977), a current component for the work during the two decades between the two papers and a future component.

Regarding the past, Meng and van Dyk do a remarkable job of summarizing, in a few pages, many crucial contributions, but there are two notable omissions, one from each side of the Atlantic. Orchard and Woodbury (1972) with their proclamation of the 'missing information principle' was certainly a milestone in the history of the EM algorithm. Also Beale and Little (1975) was an important 'predictor' of Dempster *et al.* (1977), and Beale was, in fact, the lead discussant of our paper. Of course, it is difficult to satisfy everyone's list of critical contributions, especially given editorial pressures for brevity.

Regarding the two decades between the two papers, I believe that Dempster *et al.* (1977) sparked such popularity in the EM algorithm in so many disciplines for a particular reason beyond its numerical simplicity and stability: the focus on it as a statistical, rather than purely numerical, algorithm, whose steps have straightforward and evocative data analytic interpretations. Good statistical ideas, like good scientific ideas, are powerful for organizing and expanding our creative thoughts, especially when illustrated using a range of new and old examples, as in Dempster *et al.* (1977) and Meng and van Dyk's paper. The scholarship exhibited in the present paper for these two decades is extremely impressive and leads to their key ideas by coalescing and extending results in diverse literatures. Certainly their formulation of the alternating expectation–conditional maximization (AECM) algorithm and efficient augmentation through working parameters is the current state of the art for the EM algorithm.

Before looking at the future of EM-type algorithms, a natural question is whether there is a future for maximizers or mode finders, especially considering this era's enthusiasm for iterative simulation techniques such as Markov chain Monte Carlo (MCMC) methods. My answer, even though I have preferred simulated posterior distributions to modal estimates and standard errors for many years (e.g. Rubin (1984)), is 'absolutely yes' because practically assessing convergence in distribution in an MCMC simulation is much more difficult than assessing convergence to a mode. Thus, finding regions of high density (i.e. multiple modes) by a mode finder is nearly essential before running a simulation whose results can be believed (e.g. as argued in Gelman and Rubin (1992)).

Regardng the future, 1997 has arrived in London a few weeks early, as I have already received *The EM Algorithm and Extensions* (McLachlan and Krishnan, 1997), and we now know that the future references both Meng and van Dyk's paper (albeit an earlier version) and AECM. But is this paper the final verse to the EM song? About a decade ago an extremely talented graduate student of mine was lamenting that it was sad that EM theory is so beautiful, yet had been all done. That student, who has just spoken, has vividly dispatched his own youthful naïvete. And there is still much to do!

For example, in Liu *et al.* (1996), we propose an extension of EM, called parameter-extended EM (PXEM), that improves the M-step of the EM algorithm. The basic idea is that the M-step is generally inferentially inefficient because it acts as if the imputed sufficient statistics arose from the correct value of the parameter whereas we know that they were imputed from an interim value. The discrepancy can be revealed by deviations between the imputed values of complete data statistics and their expectations under the model, and these statistics can be activated by associating new parameters with them. The resulting adjustment to the M-step is analogous to a covariance adjustment in a randomized experiment, where we can use the discrepancy between the observed difference in covariate means and its expectation, which is zero, to adjust the estimate of treatment effect. Parameter expansion estimates more parameters and adjusts by a reduction mapping that preserves the distribution of the observed data; it can, in principle, be applied to any EM-type algorithm, preserves all the desirable properties of EM-type algorithms and can only increase (or maintain) the rate of convergence.

In the multivariate t-family example, the obvious statistic to be activated through parameter expansion is the average of the weights, whose expectation is 1, and this PXEM yields exactly the same algorithm as the optimal algorithm of Meng and van Dyk. In a random effects–variance components

TABLE 6
Comparison of numbers of iterations until convergence of EM-type algorithms

σ^2	*No. of iterations to convergence of the following algorithms:*		
	EM	*Efficient EM*	*PXEM*
1/25	197	260	190
1/4	136	173	88
1/2	243	239	59
1	830	452	140
4	6657	547	140
25	12935	466	93

model, the statistics to be activated through parameter expansion include the covariance between the data and the random effects, and the resulting PXEM algorithm can be, at least in a simple example, vastly superior to the standard EM algorithm and better than an efficiently augmented EM algorithm, as shown in Table 6, comparing the number of iterations required to converge (in one simulation) as a function of the residual model variance σ^2.

The ideas of efficient augmentation and parameter expansion are complementary: the former works at the design stage in the sense that the extra maximization for selecting a working parameter takes place before running the EM algorithm, whereas the latter works at the analysis stage in the sense that the extra maximization over the expanded parameters takes place while runnning the algorithm at each M-step. When the data are optimally augmented by design there may be little or no pay-off to parameter expansion, just as when an experiment is optimally designed there may be little or no pay-off to covariance adjustment. Also, as with Meng and van Dyk's ideas for the EM algorithm, the ideas behind PXEM can be extended to speed MCMC methods.

Meng and van Dyk stimulated our developing work on PXEM and no doubt will continue to stimulate more new work by other researchers. It is with great pleasure that I propose a vote of thanks to Meng and van Dyk for their exceptionally thought-provoking contribution.

D. M. Titterington (University of Glasgow): It is a pleasure to be present at this celebration of the EM algorithm and to comment on this interesting paper. The major contributions of the paper are to introduce a new twist to the quest for methods for speeding up the algorithm and to unify some useful general ways of modifying the algorithm while preserving its appealing properties.

It would be nice to be able to set the multivariate t-problem within quite a general class for which comparable methodology exists for minimizing the effective degree of data augmentation and thereby speeding up the algorithm, and to throw light on the tantalizing observation in Section 2.2 that the optimal value of the working parameter a leads (almost) uniquely to feasible calculations. I regret that I have not been able to crack these issues, which means that most of my remarks concern more general aspects of the EM algorithm and its variations.

I wish that I could say that I have a clear recollection of the meeting at which the original paper was read. I was undoubtedly present, but I have only vague mental pictures of the event. What I recall very vividly is a sensation of remarkable revelation and coincidence, associated with the effect that the preprint had on at least three, superficially unrelated, research interests I shared at that time: part of Gordon Murray's doctoral work concerned mixture distributions, and therefore EM; also, in an almost chance discussion with Byron Morgan, an apparently *ad hoc* algorithm that he had written down for a certain incomplete data problem in contingency tables could be fitted into the EM mould (Morgan and Titterington, 1977); and an algorithm for calculating D-optimal experimental designs could also be shown to be monotonic by interpreting it as EM. In the published account of this algorithm (Silvey *et al.*, 1978) a different monotonicity argument was printed, but the elegant EM interpretation was also available.

Since then I have often been grateful to be able to add to Meng and Pedlow's (1992) impossible task of documenting EM citations, and I have become aware of other literatures where the original paper has been exploited. For instance, I would be surprised if a current version of Table 1 did not include journals, such as *Neural Computation*, from the literature on artificial neural networks. The flexibility of

many versions of neural network models is created by the inclusion of so-called hidden units, which introduce what statisticians would call latent variables, and the training of these networks constitutes an incomplete data estimation problem for which the EM algorithm provides one operational solution. Thus this literature increasingly contains items about old and new latent structure problems, many of them dealt with using the EM algorithm. On occasion, the scale of the problem threatens to defeat 'proper' EM, and certain modifications are introduced to make the calculations feasible. One approach that has been intriguing me is to replace the occasionally complex, multivariate conditional distribution in the E-step by a simpler approximating distribution. It turns out that simplicity is achieved through an independence model called the mean field approximation, and, even though the theoretical properties of the modification remain to be established, and though I am only now beginning to grasp, intuitively, why it should be successful, empirical performance is usually extremely good. Citations of the method include Ghahramani (1995), in the context of a latent profile model, and Archer and Titterington (1996) and Zhang (1992, 1993), for application to hidden Markov models. Some of Zhang's work is based on Markov random fields, and a mean field approximation is also used in the M-step to avoid problems caused by an intractable partition function; see Dunmur and Titterington (1997) for a review of this area. This literature also contains items about more familiar EM topics. For instance, Xu and Jordan (1996) revisit various aspects of the Gaussian mixture problem, making the important practical remark that, as far as future prediction is concerned, the EM algorithm is often not really all that slow, in the sense that it typically climbs high on the likelihood surface, thereby reaching well fitting models, in comparatively few iterations.

This leads me nevertheless to mention an approach to speeding up the algorithm by splitting up not just the M-step, as in the case of expectation–conditional maximization, for instance, but both steps, creating a number of 'partial E-' followed by M-steps, which likewise improve the speed. Key references are Byrne (1996) and Neal and Hinton (1993). Hudson and Larkin's (1994) ordered subsets method for image reconstruction is in the same spirit but needs a minor modification to fit the approach precisely.

Such clever variations seem to me to constitute one of the most interesting current trends in EM. Others include modifications based on Monte Carlo methods, which we shall hear about later, and revealing work on the basic framework, involving the information geometry of Amari (1995), which also underlies material in Csiszár and Tusnády (1984), Byrne (1993) and Neal and Hinton (1993), as well as special case applications to mixtures, in Hathaway (1986), and to Boltzmann machines, in Byrne (1992) and Anderson and Titterington (1995); a key observation is the interpretation of the E-step also as a maximization, and consequently of the EM algorithm as an alternating maximization algorithm of the negative of the Kullback–Leibler divergence between two probability distributions.

Altogether, the EM algorithm has had an extraordinary influence, and I am very pleased to second the vote of thanks to the authors of this further contribution.

Walter R. Gilks (Medical Research Council Biostatistics Unit, Cambridge):

Comparing EM and Markov chain Monte Carlo algorithms

The authors draw some interesting parallels between EM algorithms and the Gibbs sampler, which is a particular form of Markov chain Monte Carlo (MCMC) algorithm. Such parallels indicate the potential for cross-fertilization of ideas between these two fields, and some instances of this are referenced in Section 4. I shall focus my discussion on the comparison between EM and MCMC methods and explore further the potential for cross-fertilization.

To demonstrate the similarity between the two techniques, it is convenient to assume, in the notation of the paper, an exponential family for $p(Y_{\text{aug}}|\theta)$, where Y_{aug} is sufficient for θ. Most MCMC applications are in the Bayesian paradigm, so I shall assume a prior $p(\theta)$ for θ. Then we have the following:

(a) *EM algorithm*,

$$\begin{aligned} Y_{\text{aug}}^{(t+1)} &= E(Y_{\text{aug}}) \quad \text{under } p(Y_{\text{aug}}|Y_{\text{obs}}, \theta^{(t)}), \\ \theta^{(t+1)} &= \max\{p(\theta|Y_{\text{aug}}^{(t)})\} \quad \text{with respect to } \theta, \\ &\vdots \\ \theta^{(\infty)} &= \max\{p(\theta|Y_{\text{obs}})\} \quad \text{with respect to } \theta \end{aligned}$$

where '=' denotes assignment;

(b) *Gibbs sampler*,

$$Y_{\text{aug}}^{(t+1)} \sim p(Y_{\text{aug}} | Y_{\text{obs}}, \theta^{(t)}),$$
$$\theta^{(t+1)} \sim p(\theta | Y_{\text{aug}}^{(t)}),$$
$$\vdots$$
$$\theta^{(\infty)} \sim p(\theta | Y_{\text{obs}})$$

where '$\sim$' denotes 'sample from'.

In a Bayesian context, the EM algorithm finds the mode of the marginal posterior $p(\theta | Y_{\text{obs}})$, as noted by Dempster *et al.* (1977), and the Gibbs sampler produces samples from it. Thus the two algorithms depend on the same distributions, one deterministically and the other stochastically.

The paper is mainly concerned with rates of convergence for the EM algorithm, for which explicit formulae are provided. This is an enviable position for most MCMC practitioners, where rates of convergence are generally not calculable, and even upper bounds on rates of convergence are difficult to come by. The notable exception to this is for the Gibbs sampler in problems where the posterior distribution is multivariate Gaussian, or approximately so, for which exact (or approximate) rates of convergence have been obtained (Roberts and Sahu, 1996, 1997). For many standard problems, the MCMC algorithm is rapidly mixing (i.e. it converges quickly), and the computation time is of the order of a few minutes. However, MCMC methods are increasingly being used in non-standard, complex, applications, where mixing can be very slow, the computations typically taking days to complete. To improve mixing, MCMC practitioners have become adept at devising auxiliary variables (data augmentation) and reparameterization schemes.

Thus I was particularly interested in the authors' data augmentation–reparameterization scheme for the t-model in Sections 2.2 and 2.3. Representing a t-distribution as a scale mixture of normals is a familiar technique, but the proposed reparameterization in equation (2.5) was quite unexpected. The conjecture in Section 4, that the reparameterization might also work well for the Gibbs sampler, was irresistible, so I performed my own simulation study.

Simulation study for t-model

For several values of ν, I generated 100 independent realizations from a $t_1(0, 1, \nu)$ distribution and modelled them with a $t_1(\mu, \tau^{-1}, \nu)$ distribution, where $\tau = \Sigma^{-1}$, the inverse of the scale parameter. With the reparameterization in equation (2.5), the full conditional distributions $p_{a,\nu}(\mu|\tau, \{q_i, i = 1, \ldots, 100\})$ and $p_{a,\nu}(q_i|\tau, \mu)$ were straightforward to sample from, being Gaussian and gamma respectively. However, the full conditional for τ, $p_{a,\nu}(\tau|\mu, \{q_i, i = 1, \ldots, 100\})$, was quite unpleasant, requiring adaptive rejection Metropolis sampling (ARMS) (Gilks *et al.*, 1995).

For each value of ν and several values of a, I ran 10000 iterations of the Gibbs sampler. Each run took approximately 30 s on a Sun SPARC station. Examining the output revealed that the marginal sequence for τ was approximately an AR(1) process, so the mixing rate in the sequence could be adequately represented by its lag 1 autocorrelation ρ_1. Table 7 reports the estimated ρ_1 for each value of a and ν.

Thus, setting $a = 1/(1 + \nu)$ gives the optimal rate of convergence, just as for the EM algorithm. For each ν, the improvement over the usual parameterization $a = 0$ is about 35%, which is not quite as good as reported for the EM algorithm in Fig. 2(b). Setting $a = 1/(1 + \nu)$ would therefore allow a shorter chain to be run (for the same precision in the estimate of the posterior mean of τ), with a saving in

TABLE 7
Missing rates for the Gibbs sampler applied to the t_1-model

ν	*Estimated ρ_1 for the following values of a:*								
	0.0	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*	*0.6*	*0.7*	*0.8*
1	0.76	0.70	0.64	0.58	0.54	0.49	0.52	0.56	0.63
2	0.59	0.49	0.42	0.39	0.39	0.43	0.51	0.58	0.66
4	0.43	0.33	0.28	0.32	0.43	0.53	0.62	0.69	0.76
9	0.22	0.14	0.22	0.39	0.56	0.66	0.75	0.80	0.83

computation time of about 12 s. However, with $a = 0$, the full conditional for τ reduces to a gamma distribution. Exploiting this reduced the computation time for $a = 0$ by about 10 s. So the computational advantage of reparameterization is negligible. Moreover, the additional work in coding the algorithm for general a (about a day!) makes the reparameterization quite unattractive.

An additional complication is that the full conditional distribution for τ fails to be demonstrably log-concave just below $a = 1/(1 + \nu)$. Log-concavity in full conditionals is desirable since pure Gibbs sampling is difficult without it; the ARMS method includes a Hastings–Metropolis step to account for non-log-concavity. Log-concavity also provides reassurance about unimodality of the posterior. For the t-model in d dimensions, the optimal EM algorithm sets $a = 1/(d + \nu)$. Again, the full conditional distribution for precision matrix τ ceases to be demonstrably log-concave just below this optimal value of a.

Fig. 3 shows much greater benefits in reparameterizing the EM algorithm for the t-model in higher dimensions, nearly tenfold for $d = 10$. If such improvements carry over to the Gibbs sampler, as the present results suggest they might, then reparameterization might be worth considering. Nevertheless, computational savings must be balanced against increased human effort. In situations much more complex than those presented in the paper, where mixing is extremely slow and each iteration takes several seconds, reparameterization has much greater potential. However, the paper gives little insight into how reparameterization should be done in applications other than the two discussed.

Speed of convergence is clearly of much greater importance in the MCMC than in the EM algorithm. The suggestion in Section 4 that efficient Gibbs sampler implementations might be sought by first seeking efficient EM implementations seems extremely valuable and worthy of further study.

Implementing EM algorithm by using Markov chain Monte Carlo method

The authors do not discuss ways in which the MCMC method can be used to implement the EM algorithm. It can be used both in the E-step and in the M-step.

At the E-step, the required expectation may be difficult to obtain in closed form. Wei and Tanner (1990) suggested independent sampling from $p(Y_{\text{aug}}|Y_{\text{obs}}, \theta^{(t)})$ to compute a Monte Carlo estimate of the expectation (the MCEM algorithm). When independent sampling is impractical, correlated samples can be generated by using the MCMC algorithm. Thus a full MCMC run would be required within each EM iteration. This method has been used in analysing complex pedigrees (Guo and Thompson, 1994). This algorithm is different from the stochastic EM algorithm proposed by Celeux and Diebolt (1985), in which only one iteration of the MCMC algorithm is performed within each EM iterate. The stochastic EM algorithm has the disadvantage of producing only approximate maximum likelihood estimates.

At the M-step, in situations where the log-likelihood $L(\theta; Y_{\text{aug}}^{(t)})$ may be multimodal, simulated annealing could be used to avoid being trapped in local modes. Simulated annealing is essentially an MCMC method where, at each iteration, the stationary distribution $p(\theta|Y_{\text{aug}}^{(t)})$ is made progressively more concentrated around its mode, by raising it to an increasingly positive power. The various conditional schemes suggested in the paper find a natural home within this framework.

Jean Diebolt (Université de Grenoble): Meng and van Dyk introduce a new version of the EM algorithm, the alternating expectation–conditional maximization (AECM) algorithm. Theorem 4 provides an expression for the Jacobian matrix $DM^{\text{AECM}}(\theta^*)$ at any limiting point θ^* of the AECM scheme. It shows that this matrix is 'smaller' than the corresponding matrix for the EM algorithm, implying that the AECM algorithm is faster when the current estimate is already close to θ^*. However, since the observed data (OD) log-likelihood function (LF) is non-decreasing, nothing prevents it from converging to the nearest local maximum or from being trapped on a plateau of the log-likelihood surface. In examples 1 and 2 the ODLF is unimodal. Possible multimodality of the ODLF and related dependence on starting values were not considered.

Weak identifiability

In many situations of interest the model with missing data is *weakly identifiable*, i.e. the ODLF exhibits several local maxima of comparable magnitude (up to sampling fluctuations) even for comparatively large sample sizes n. This is true, for example, for finite mixtures. Results on the existence and uniqueness of a strongly consistent sequence of maximum likelihood estimates among these local maxima (e.g. Redner and Walker (1984)) do not help in deciding which is better. This is a difficult problem, as stressed by Robert (1992). (See Robert and Soubiran (1993) for the prior feed-back approach.) To my knowledge, no satisfactory solution to this problem is available from a frequentist point of view.

The EM algorithm and more advanced versions such as AECM do not consider this problem. They converge to the closest local maximum (i.e. a stable fixed point of M^{EM} or M^{AECM} of the ODLF, if not trapped first.

Stochastic versions of EM algorithm

Stochastic versions of the EM algorithm have been introduced to address this problem. Some, such as stochastic EM (SEM) or stochastic approximation EM (SAEM) and its variants, can be used in contexts where the EM algorithm is intractable.

Wei and Tanner (1990) introduced Monte Carlo EM (MCEM) when the computation of $Q(\theta'|\theta)$ is intractable. However, MCEM does not address weak identifiability and can lead to impractical procedures.

Lavielle and Moulines (1997) studied the stochastic approximation version SAEM and a simulated annealing variant. They established the almost sure convergence of the SAEM algorithm to a local maximum and the convergence in distribution of the annealing variant to the global maximum of the ODLF.

The SEM algorithm (e.g. Celeux and Diebolt (1985, 1988)) incorporates an S-step which simulates a realization $Z^{(t)}$ of the missing data set from the posterior density $k(Z|Y_{obs}, \theta^{(t)})$ based on the current estimate, which is then updated by maximizing the LF of the restored data set $(Y_{obs}, Z^{(t)})$. Since there is no need to compute $Q(\theta'|\theta^{(t)})$, human effort is minimal. If necessary, the S-step can be completed through Gibbs sampling or Hastings–Metropolis sampling. The code for the SEM algorithm provides a Monte Carlo approximation of the OD information through Louis's (1982) formula (Diebolt and Ip, 1996).

Theoretical results on ergodicity and asymptotic behaviour as $n \to \infty$ of the invariant distribution $\Psi(d\theta|Y_{obs})$ of the homogeneous Markov chain generated by the SEM algorithm have been obtained, e.g. in Diebolt and Celeux (1993) and Ip (1994), under restrictive conditions. The SEM algorithm explores the global behaviour of the discrete time dynamical system generated by M^{EM} by adding a random perturbation ϵ at each iteration:

$$\theta^{(t+1)} = M^{EM}(\theta^{(t)}) + \epsilon_{t+1}.$$

In many cases of interest, ϵ_{t+1} is approximately normal with mean 0 and variance

$$\Sigma^{(t+1)} \approx n^{-1} I_{aug}^{-1}(\theta^{(t)})\{I - I_{obs}(\theta^{(t)}) I_{aug}^{-1}(\theta^{(t)})\}.$$

The stationary regime of SEM provides a smooth picture of the repartition of the most stable fixed points of M^{EM}. Furthermore, the mean $\bar{\theta}_{SEM}$ of $\Psi(d\theta|Y_{obs})$ provides a simple alternative to the maximum likelihood estimate.

Also, Celeux and Diebolt (1992) and Biscarat (1994) have established that a version starting from SEM at the beginning to EM at the end converges almost surely to a local maximum of the ODLF, thus avoiding traps.

Conclusion

The idea of 'augmenting less' by tuning artificial parameters is very attractive and stimulating. It would be interesting to experiment with the AECM algorithm in weakly identifiable settings such as finite mixtures, where the EM algorithm can be very slow. How should optimal values of these working parameters be determined in such a case? Also, could the AECM algorithm substantially improve on the EM algorithm in situations where the EM algorithm is barely tractable such as in hidden Markov models (e.g. Qian and Titterington (1992)), where Gibbsian versions of the SEM algorithm can easily be implemented as in Robert *et al.* (1993)?

The conditioning and change-of-variables ideas of AECM can be adapted to SEM and SAEM. We hope that they will improve the efficiency of these stochastic versions of the EM algorithm without sacrificing their conceptual and implementational simplicity, and their ability to cope with weakly identifiable situations.

The following contributions continued the discussion.

Murray Aitkin (University of Newcastle): This paper performs the valuable service of opening our eyes wider on the formulation of EM algorithms. It has long been clear that there may be several

different ways of implementing the EM algorithm for any given problem. For example, in the binary probit factor analysis model of Bock and Aitkin (1981), the complete data model could be formulated as a normal factor model with unobserved response variables, or a set of independent probit regression models with unobserved explanatory variables. Software for the M-step was much more flexible for the second implementation than for the first, so we used the second.

The paper shows that we can go further than this, if we are lucky with the model, and find an *optimal* formulation of the missing data model which gives the fastest convergence of the algorithm. Variance component models provide a simple illustration.

Consider the simple unbalanced one-way normal variance component model. The usual formulation is

$$y_{ij}|a_i \sim N(\mu + a_i, \sigma^2), \qquad a_i \sim N(0, \sigma_A^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, r, \quad \sum n_i = N.$$

It is easily shown that the complete data expected information matrix is

$$I_1 = \begin{pmatrix} N/\sigma^2 & 0 & 0 \\ 0 & 2N/\sigma^2 & 0 \\ 0 & 0 & 2r/\sigma_A^2 \end{pmatrix}.$$

The rate of convergence is governed by the size of σ_A^2: convergence will be fastest when $2r/\sigma_A^2$ is as small as possible and will be very slow when σ_A^2 is near 0, as is well known.

We can formulate the complete data model differently, however:

$$y_{ij}|a_i \sim N(\mu + \sigma_A a_i, \sigma^2), \qquad a_i \sim N(0, 1).$$

Now the complete data expected information is

$$I_2 = \begin{pmatrix} N/\sigma^2 & 0 & 0 \\ 0 & 2N/\sigma^2 & 0 \\ 0 & 0 & N/\sigma^2 \end{pmatrix}$$

since now σ_A is essentially a regression coefficient. Surprisingly, the rate of convergence does not depend on σ_A in this implementation. The relative rates of convergence depend on the ratio

$$\frac{N/\sigma^2}{2r/\sigma_A^2} = \bar{n}\theta/2$$

where θ is the variance component ratio and $\bar{n}$ is the average class sample size. Method 2 is faster if $\theta < 2/\bar{n}$.

Now can we formulate an optimal balance between these algorithms? Consider the formulation

$$y_{ij}|a_i \sim N(\mu + \sigma_A^\lambda a_i, \sigma^2), \qquad a_i \sim N(0, \sigma_A^{2(1-\lambda)}).$$

This preserves the marginal distribution of the y_{ij} and the intraclass correlation, whatever the value of λ in the range [0, 1].

The expected information in this case is

$$I_3 = \begin{pmatrix} N/\sigma^2 & 0 & 0 \\ 0 & 2N/\sigma^2 & 0 \\ 0 & 0 & N\lambda^2/\sigma^2 + 2r(1-\lambda)^2/\sigma_A^2 \end{pmatrix}.$$

The previous cases correspond to $\lambda = 0$ and $\lambda = 1$. The information is minimized when $\lambda = 1/(1 + \bar{n}\theta/2)$ and is then $N/\sigma^2(1 + \bar{n}\theta/2)$.

TABLE 8
Complete data expected information for three models

θ	I_1	I_2	I_3
0	∞	30	30
0.1	100	30	23.1
0.2	50	30	18.8
0.3	33.3	30	15.8
0.4	25	30	13.6
0.5	20	30	12
1	10	30	5

To illustrate, consider an example with $r = 5$, $\bar{n} = 6$ and $\sigma^2 = 1$, for various values of θ (Table 8). The optimal algorithm is as fast as method 2 when $\sigma_A^2 = 0$ and faster than either method when $\sigma_A^2 > 0$. Of course it cannot be implemented optimally without knowledge of θ, but still a rough estimate of θ may provide much faster convergence than either standard form. Alternatively, $\hat{\theta}$ may be used to update λ at each iteration.

Cedric A. B. Smith (University College London): In Naples in 1953 it struck me that the EM algorithm, to be discovered by Dempster *et al.* (1977), solved many genetical estimation problems (Ceppellini *et al.*, 1955; Smith, 1957). Sometimes there are snags. A model for the estimation of inbreeding from ABO blood group frequencies is shown in Table 9.

Here, a, b and c are the frequencies of the A-, B- and O-genes, and f and g are coefficients of inbreeding and outbreeding, so that $a + b + c = 1 = f + g$. Using the EM algorithm, my collaborator, Robert Thomson, found that even after hundreds of iterations there was little sign of convergence. Unexpectedly, it happens that the maximum likelihood equations have two explicit solutions shown in Table 10 (Schull and Ito, 1969).

The likelihood surface has a ridge, values of likelihood falling inappreciably from the maximum at points on the line joining these two values, resulting in slow convergence. Another problem was found in the investigation of the inheritance of tuberous sclerosis, where sometimes it is due to a gene on chromosome 9, sometimes not (Smith and Stephens, 1996). The EM algorithm converged rapidly to a boundary maximum, so that the usual formulae for standard error failed.

There may occasionally be quicker algorithms. In 1944 Tweedie considered a genetic estimation with maximum likelihood equation

TABLE 9
Data on ABO blood group frequencies

Blood group	*Expected frequency*	*Observed no.*
A	$ag(a + 2c) + af$	81
B	$bg(b + 2c) + bf$	32
AB	$2abg$	17
O	$c(cg + f)$	51

TABLE 10
Maximum likelihood estimates

a	b	c	f	g
0.340	0.154	0.506	0.101	0.899
0.305	0.138	0.556	−0.112	1.112

$$\frac{130}{R}-\frac{515}{2-R}-\frac{161}{1-R}+\frac{461}{1+R}=0. \tag{1}$$

He thought that it would be easily solved if we just inverted the fractions, giving a linear equation:

$$\frac{R}{130}-\frac{2-R}{515}-\frac{1-R}{161}+\frac{1+R}{461}=0. \tag{2}$$

In fact, the solution of equation (1) is $R = 0.439$ and, of equation (2), $R = 0.440$, negligibly different. For an explanation, see Smith (1969, 1997).

John Hinde (University of Exeter): Consider a generalized linear model (GLM) with an additional normal random effect in the linear predictor. For observed responses y_i with covariates $\mathbf{x}_i$, $i = 1, \ldots, n$, the linear predictor is

$$\eta_i = \boldsymbol{\beta}'\mathbf{x}_i + \tau z_i$$

where the $\{z_i\}$ are independently $N(0, 1)$ distributed. Maximum likelihood fitting of this model using the EM algorithm is described in Anderson and Hinde (1988). The E-step is evaluated using Gaussian quadrature by

$$Q^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \approx \sum_{j=1}^{K}\sum_{i=1}^{n} w_{ij}^{(t)} \log f(y_i|z_j, \boldsymbol{\theta}) + \sum_{j=1}^{K}\sum_{i=1}^{n} w_{ij}^{(t)} \log \phi(z_j)$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \tau^2)$, $\{z_j\}$ are the quadrature points for an $N(0, 1)$ distribution and $\{w_{ij}^{(t)}\}$ are the weights based on the conditional density of y_i given z_j and the current estimate $\boldsymbol{\theta}^{(t)}$. The M-step involves only the first term and requires a standard GLM fit to expanded data of length nK with $\mathbf{x}_i$ and z_j as explanatory variables and weights $w_{ij}^{(t)}$.

An alternative formulation has

$$\eta_i = \boldsymbol{\beta}'\mathbf{x}_i + u_i$$

where the $\{u_i\}$ are now independently $N(0, \tau^2)$ distributed. The EM algorithm now has E-step

$$Q^{(t+1)}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \approx \sum_{j=1}^{K}\sum_{i=1}^{n} w_{ij}^{(t)} \log f(y_i|u_j, \boldsymbol{\beta}) + \sum_{j=1}^{K}\sum_{i=1}^{n} w_{ij}^{(t)} \log\left\{\frac{1}{\tau}\phi\left(\frac{u_j}{\tau}\right)\right\}$$

where the $\{u_j\}$ are now quadrature points for an $N(0, (\tau^{(t)})^2)$ distribution and $\{w_{ij}^{(t)}\}$ are again conditional weights. The M-step splits into the estimation of $\boldsymbol{\beta}$ by a GLM fit to expanded data with weights $w_{ij}^{(t)}$, u_j as a fixed offset and the $\mathbf{x}_i$ as explanatory variables, whereas the estimates of τ^2 is obtained from the second term as $\Sigma_j \Sigma_i w_{ij}^{(t)} u_j^2/n$. The computational effort at each iteration is similar for both formulations; however, the convergence rates can differ.

McCulloch (1994) presented a probit–normal model for data on the survival of rats in a treatment and a control group, with litter as a single nested random effect. Writing y_{ij} and n_{ij} for the numbers of survivors and totals in each group–litter combination, McCulloch's model corresponds to $y_{ij} \sim \text{bin}(n_{ij}, p_{ij})$, $i = 1, 2, j = 1, 2, \ldots, 16$, where

$$\Phi^{-1}(p_{ij}) = \mu_i + u_{ij}$$

with $u_{ij} \sim N(0, \tau_i^2)$. For the control group ($i = 1$), using 20-point quadrature, McCulloch noted that the EM algorithm required over 200 iterations. With the z-formulation, convergence requires only 30 iterations.

The relative performance of the two formulations is governed by the size of the random effect variance. For single-parameter exponential families when τ^2 is small the z-formulation is faster, whereas, for large values of τ^2, the u-formulation is faster. In the simple one-way normal variance components model, where closed form estimates can be obtained for both formulations, the relative speed depends on the ratio of τ^2 to σ^2, the residual variance—this can also be seen from I_{aug}.

One way to find an optimal EM algorithm would be to write

$$\eta_i = \beta' \mathbf{x}_i + \tau^a v_i$$

where the $\{v_i\}$ are independently $N(0, \tau^{2(1-a)})$ distributed. This includes the above formulations for $a = 1$ and $a = 0$. However, the optimal choice of a will depend on the unknown parameters and for general values of a we no longer have the simple M-steps as described above. Could an alternating expectation–conditional maximization algorithm be applied in this case?

John T. Kent (University of Leeds): It is a pleasure to see that there is still so much life left in the EM algorithm after 20 years of activity. The authors discuss two algorithms introduced in Kent *et al.* (1994) on the estimation of the mean vector μ and scatter matrix Σ of the p-dimensional multivariate t-distribution with $\nu > 0$ degrees of freedom. To understand our motivation behind these two algorithms it is useful to start with the multivariate Cauchy distribution, for which there are two natural incomplete data representations, noted as early as Sundberg (1974):

$$Y = \mu + Z/|V|, \qquad Z \sim N_p(0, \Sigma) \text{ independent of } V \sim N_1(0, 1),$$

$$Y = W/V, \qquad \begin{pmatrix} W \\ V \end{pmatrix} \sim N_{p+1}\left\{0, \lambda \begin{pmatrix} \Sigma + \mu\mu^{\mathrm{T}} & \mu \\ \mu^{\mathrm{T}} & 1 \end{pmatrix}\right\}.$$

These two representations, a location–scatter problem in p dimensions and a scatter-only problem in $p+1$ dimensions respectively, generate the two EM algorithms in the paper. The scatter-only representation is also useful for proving uniqueness in this and related problems (Kent and Tyler, 1991).

The first representation extends easily to all $\nu > 0$, thus guaranteeing that the first algorithm is always an EM algorithm. The second representation can be extended in a somewhat artificial way to $\nu > 1$, and the second algorithm arises as a *modification* of an EM algorithm. In spite of our best efforts in Kent *et al.* (1994), we were unable to show that it was a proper EM algorithm. Thus we are very impressed by the authors' success for all $\nu > 0$, especially for $0 < \nu < 1$.

An intriguing feature of these two EM algorithms is that the second algorithm converges more quickly. The authors present some empirical evidence and a proof that the second algorithm converges more quickly near the maximum likelihood estimate for any data set. Arslan *et al.* (1995), section 5, gave explicit rates of convergence for the large sample case, namely

$$r_1 = (p+2)/(p+\nu+2),$$
$$r_2 = 2/(p+\nu+2),$$

where smaller values of r correspond to faster convergence.

Finally, let me ask a general question about the use of 'working parameters'. The use of $|\Sigma|^{-a/2}$ for the multivariate t seems to be an inspired guess which depends heavily on special properties of this distribution. But it is unclear to me how routine it will be to apply similar ideas in other incomplete data problems.

David E. Tyler (Rutgers University, Piscataway):

A more general framework for the EM algorithm?

Iterative reweighting algorithms for multivariate M-estimates of location and scatter have been used since the estimates were first introduced by Maronna (1976). These algorithms are similar to equations (2.3) and (2.4), but with the weight function

$$w(d) = \frac{\nu + p}{\nu + d}$$

replaced by a general weight function $w(d)$. Little is known about the theoretical convergence properties of these algorithms. Also, the iterative reweighting algorithms tend to converge especially slowly for 'bad' data sets and ironically the possibility of encountering such a set is the motivation behind using a robust statistic.

We can obtain some convergence results. If the M-estimate corresponds to the maximum likelihood estimate for some scale mixture of normals, then the iterative reweighting algorithm is an EM algorithm and so we know that it converges monotonically (Rubin, 1983). A more general setting is needed though since flexibility in obtaining desirable robustness properties is important in constructing M-estimates. If we hold the location μ fixed and consider the iterative reweighting algorithm for scatter only, then Kent and Tyler (1991) showed that this algorithm always converges regardless of the initial value whenever the weight function is non-negative and the influence function for scatter is non-decreasing. Furthermore, Maronna (1976) showed that for a sufficiently large initial value the convergence is monotone in Σ.

An important issue is that the multivariate maximum likelihood estimates, as well as the multivariate M-estimates for which the influence function for scatter is non-decreasing, have limited robustness appeal. In particular, their breakdown point can never be greater than $1/(p+1)$. For the maximum likelihood estimates based on the t-distribution with ν degrees of freedom the breakdown point is $1/(p+\nu)$ for $\nu \geqslant 1$ and $\nu/(p+\nu)$ for $0 < \nu < 1$. Thus the best breakdown point of $1/(p+1)$ occurs at the Cauchy maximum likelihood estimate.

This has led to a consideration of multivariate M-estimates whose influence functions redescend to 0; these never correspond to maximum likelihood estimates. Moreover, there are numerous solutions to the corresponding M-estimating equations of which only one specific solution may be of interest; see Lopuhaä (1989) and Kent and Tyler (1996). Thus, no easily computed multivariate statistic can be highly robust.

This does not mean that the simple reweighting algorithm has no future role in robust statistics. Ruppert (1992) has proposed a reasonably fast and accurate algorithm which uses a random subsampling algorithm coupled with an important local improvement step based on a reweighting algorithm. The local improvement step checks that the reweighting algorithm improves the objective function for the S-estimator. Another good reference is Woodruff and Rocke (1994). Some convergence results are given by Tatsuoka (1996). Finally, there are valid arguments for looking at all possible solutions to the M-estimating equations (see for example Morganthaler (1990)), and the simple reweighting algorithms are good for finding at least some solutions.

Paul Damien (University of Michigan, Ann Arbor) **and Stephen Walker** (Imperial College of Science, Technology and Medicine, London): With respect to parameter estimation, in addition to devising algorithms that are 'simple, stable and fast', as the authors correctly identify from a maximization perspective, it is also essential that they be *general*. With this in mind and from a Bayesian posterior sampling perspective, after introducing auxiliary variables, a Gibbs sampler can be constructed so that *all* the full conditional distributions are uniform.

Theorem 1 (Damien and Walker, 1996a). If

$$f(x) \propto \prod_{l=1}^{L} g_l(x),$$

where the g_l are non-negative invertible functions (not necessarily densities)—i.e. if $g_l(x) > u$, then it is possible to obtain the set $A_l(u) = \{x: g_l(x) > u\}$—then it is possible to implement a Gibbs sampler in which all the full conditionals will be uniform densities.

Such a class of densities is large and numerous examples have now been presented: Cumbus *et al.* (1996); Damien and Walker (1996a, b); Walker and Damien (1996a, b).

The value of the approach is that the use of rejection-based methods, sampling–resampling algorithms and Metropolis–Hastings sampling are bypassed in many contexts: these methods require identifying dominating densities and calculating supremums and acceptance rates, which may be difficult to obtain. In addition, the coding of a Gibbs sampler comprising uniform densities alone is trivial.

The following contributions were received in writing after the meeting.

Didier Chauveau (Université de Marne la Vallée, Noisy-le-Grand): Meng and van Dyk assume that we should focus essentially on two aspects when looking for alternatives to the EM algorithm: speed of convergence and ease of implementation (*simplicity*). An additional desirable aspect related to *stability* is monotone convergence in likelihood.

Simplicity

The EM algorithm itself may not satisfy the second criterion (simplicity), in common situations leading to intractable M-steps. This arises for example when fitting finite mixtures of non-exponential-type distributions with right-censored data. Then $Q(\theta|\theta^{(t)}) = \mathbb{E}[L(\theta|\mathbf{y}_{\text{aug}})|\mathbf{y}_{\text{obs}}, \theta^{(t)}]$ is not available in closed form, and finding $\arg\max_\theta\{Q(\theta|\theta^{(t)})\}$ leads to difficulties in implementation. An illustration of this takes place in actual industrial situations (Chauveau, 1992), with censored mixtures of Weibull densities $w(\cdot|\beta, \eta)$, where the M-step requires the computation of, formally,

$$(\beta, \eta)^{(t+1)} = \arg\max_{(\beta,\eta)} [\varphi\{\mathbb{E}[\log\, w(\cdot|\beta, \eta)|\mathbf{y}_{\text{obs}}, (\beta, \eta)^{(t)}]\}]$$

where φ is some function involving other terms depending on the mixture. Maximization in the Weibull case requires two nested numerical iterative methods in each EM iteration: one for finding $(\beta, \eta)^{(t+1)}$ (typically Newton–Raphson type) and one for each evaluation of the integral term inside the maximization algorithm.

Other industrial experiments, producing data through several intricate censoring schemes including left and right censoring, grouping and truncation (Gouno, 1996), or other situations (e.g. Lavielle and Moulines (1997)) result in M-steps whose implementation is beyond human and computer resources.

The alternating expectation–conditional maximization algorithm and model reduction ideas presented by the authors are interesting for handling situations where the M-step requires iterations, but the CM-steps can have analytic solutions. In the Weibull situation above, however, model reduction does not help: conditionally on η, the maximum likelihood estimate of β is still non-analytic even in a simple mixture situation. Moreover, with censored data, the integral term above remains in non-closed form, even with constrained maximization.

Stochastic alternatives

Stochastic versions of EM (SEM, Celeux and Diebolt (1985, 1992)) which restore the missing data $\mathbf{z}^{(t+1)} \sim k(\cdot|\mathbf{y}_{\text{obs}}, \theta^{(t)})$, were designed primarily to overcome the dependence of the EM algorithm on its starting point and convergence to saddlepoints of $L(\theta|\mathbf{y}_{\text{obs}})$. In situations that are intractable for the EM algorithm, the SEM algorithm also provides convenient alternatives, since the computation of $Q(\theta|\theta^{(t)})$ is not required, and the M-step reduces to easy-to-implement maximum likelihood estimation on complete data. In the Weibull example, the non-closed-form integral term simply disappears for suitable versions of the SEM algorithm (Chauveau, 1995).

For the SEM alternative, the saving in computing time is not at the expense of greater human effort. Moreover, the SEM algorithm may exhibit a higher local maximum of $L(\theta|\mathbf{y}_{\text{obs}})$ (Celeux *et al.*, 1996).

David Draper (University of Bath): I have two comments on this interesting paper.

Firstly, scale and shape are confounded in the t-family; for example, in the univariate location model

$$Y_i = \theta + \sigma e_i, \qquad e_i \overset{\text{IID}}{\sim} t_\nu,$$

the variance of Y_i is $\sigma^2\nu/(\nu - 2)$ for $\nu > 2$. Does the authors' use of the optimal working parameter a_{opt} simply correspond to a reparameterization that minimizes the posterior correlation between the transformed scale and shape parameters?

Secondly, a possible approach, alternative to that advocated in this paper, to the fitting of models like those in Sections 2.2 and 3.4 is based on Metropolis sampling and simulated annealing (SA). After transforming all parameters to live on $(-\infty, \infty)$, a reasonably generic Metropolis sampler — based on a multivariate normal proposal distribution $N_p(\mu, \Sigma)$, centred at the last value accepted and using for Σ a multiple k of an estimate of the posterior covariance matrix (either based on the negative inverse Hessian of the log-posterior, determined symbolically or numerically, evaluated at an estimate of the posterior mean or global posterior mode, or estimated adaptively from the output of the sampler) — takes about 50 lines of S code, and no more than 5–10 additional lines are needed to embed this sampler in an SA algorithm with, for example, a geometric cooling schedule. k can be chosen to optimize the sampler's convergence properties, e.g. by achieving a Metropolis acceptance rate of (say) 30–70%. Running the SA algorithm with a variety of starting values parallels the EM activity of searching for multiple modes, and fixing the temperature parameter at 1 provides Markov chain Monte Carlo (MCMC) estimates of the complete marginal posterior distributions for the parameters and any interesting functions of them (see Draper and Cheal (1997) for an example of this strategy applied to selection models in causal inference).

It is arguable that strategies like the approach sketched here — together with careful use of MCMC diagnostics — may be used to provide complete solutions to Bayesian computational problems, with roughly the same human time needed with the EM approach just to find posterior modes, and often with only a modest increase in computer time. Yes, there are applications in which MCMC extraction of the full posterior can be computationally prohibitive, but what do you do in such problems with the EM approach, after you have found the modes, to obtain the whole posterior as accurately and more quickly? To put the question sharply, in the interests of a vigorous discussion, what place has the EM algorithm in an MCMC world?

Jerome A. Dupuis (Université Paul Sabatier, Toulouse): I would like to discuss the paper in relation to the original aim of the EM algorithm of Dempster *et al.* (1977), namely to provide the maximum likelihood (ML) estimator in the presence of missing data. To do this, we focus on discrete time longitudinal data (such as imperfect medical follow-up or capture–recapture experiments) which constitute an important class of missing data model. The data that we consider are made of n independent and identically distributed realizations of a non-homogeneous Markov chain $\{y_t: t = 1, \tau\}$, with $k \geqslant 2$ states, and whose transition probabilities are denoted by θ. In this missing data set-up, numerical methods have important difficulties in providing the ML estimate of θ (Dupuis, 1995). The computational difficulties, due to the complicated form of the likelihood, stress the necessity of implementing a specific algorithm such as the EM algorithm or versions of it. Although the complete likelihood $L(\theta|y, z)$, where y are the observed data and z the missing data, has an explicit form, the EM algorithm is not very attractive here since the maximization of $E[L(\theta|y, z)]$ is laborious to implement. The alternating expectation–conditional maximization (AECM) algorithm of Meng and van Dyk could be tried to overcome these difficulties and to speed up the EM algorithm, but we think that the AECM algorithm would be cumbersome to implement. In contrast, the stochastic version of the EM algorithm of Celeux and Diebolt (1985), which also significantly speeds it up, is a very attractive alternative because of its great simplicity. First the maximization step is immediate here because it is carried out directly on $L(\theta|y, z)$ instead of on its expectation as in the AECM algorithm. The stochastic step is also immediate to implement since the conditional distribution of z given θ and y has a simple form. The SEM estimator is the mean of the stationary distribution of the Markov chain $(\theta^{(l)})$ which turns out to be ergodic (Dupuis, 1995), Moreover the SEM algorithm should significantly mitigate the main drawbacks shared by the EM and AECM algorithms: the strong dependence on the starting value and the possible convergence to spurious local maxima (Diebolt and Ip, 1996). We have compared the SEM estimator and the Bayesian estimator for non-informative prior distributions. For instance, with 25% missing data, $k = 2$, $\tau = 5$ and $n = 13$, the differences between the two estimators do not differ more than 0.01 when $n = 13$. This highlights the facts that the SEM algorithm can be viewed as a non-informative version of the Gibbs sampling algorithm and that, conversely, non-informative Bayesian procedures provide an alternative for obtaining the ML estimator.

Jeffrey Fessler (University of Michigan, Ann Arbor): In the medical imaging field, an EM algorithm has finally reached the 'pop charts'. Companies that make instruments for single-photon emission computed tomography (SPECT) now sell EM algorithm software for image reconstruction. (Interestingly, it was the non-uniform attenuation problem in SPECT that drove this evolution, not the considerations of noise or spatial resolution addressed in many EM-related papers for positron emission tomography.) Sadly, it is the 'golden oldies' version of that EM algorithm (Shepp and Vardi, 1982) that is commercially available — without regularization or acceleration. (In fact, the imaging community parochially calls it 'the' EM algorithm!) Our efforts to 'jazz up' EM algorithms with penalty functions and faster convergence (Fessler *et al.*, 1993; Fessler and Hero, 1995) have yet to make the commercial 'hit parade'.

Within part of the medical imaging community, the continuing prevalence of classical EM over contemporary faster renditions is due to a lack of acceptance of methods for regularization. Without regularization, methods for accelerating EM algorithms for tomography are of limited use, because they only provide faster convergence to poor images (since the maximum likelihood estimate has unacceptably high variance). One problem with regularization has been that conventional penalty functions result in non-uniform spatial resolution, with the poorest resolution in regions of highest intensity. This property is counter-intuitive and undesirable, and we have partially addressed it in Fessler and Rogers (1996). Further work in understanding and improving the penalty functions will be needed before

regularized image reconstruction methods and the accompanying fast algorithms will see widespread use in medical imaging.

In classical missing data problems in statistics, there is often exactly one natural choice for the augmented data, so the EM algorithm is appealing. In contrast, in the image reconstruction problem described in Fessler and Hero (1995) and summarized in Section 3.5, the augmented data have no physical interpretation. As noted in Fessler and Hero (1995) one can derive the space alternating generalized EM (SAGE) algorithm for image reconstruction from a non-statistical perspective by using only the concavity of the log-likelihood and the convexity inequality. In some respects this derivation is considerably simpler than the EM-based approach of Section 3.5, since it avoids the artificial augmented data. We refer the reader to Fessler *et al.* (1997) for details about this alternative derivation, which encompasses a large class of problems.

Finally, the reader should be aware that the convergence results in Section 3.3 do not apply to the image reconstruction problem in Section 3.5 owing to the non-negativity constraint. A general proof in Fessler and Hero (1995) establishes convergence of the SAGE algorithm for image reconstruction.

Andrew Gelman (Columbia University, New York): This paper and Meng and van Dyk (1997) have changed how I think about computation for mixed effects regression models, the simplest of which have the form

$$y \sim N(X\beta + W\alpha, \Sigma), \qquad \alpha \sim N(0, \tau^2 I), \tag{3}$$

where y, α and β are vectors. The joint maximum likelihood (ML) estimate of (α, β, τ) has poor properties if the dimension of α is high, and the EM algorithm can be used to obtain the marginal ML estimate of (β, τ); see Laird and Ware (1982). When the estimate of τ is near 0, the parameters τ and $\|\alpha\|$ tend to be highly correlated, with the unfortunate result that the EM algorithm (and the related Gibbs sampler) can move very slowly.

This paper suggests breaking the correlation by using the parameterization

$$y \sim N(X\beta + \tau^a W\gamma, \Sigma), \qquad \gamma \sim N(0, \tau^{2-2a} I),$$

where the vector $\gamma = \tau^{-a}\alpha$ is treated as 'missing data' in the EM algorithm and a is a scalar 'working parameter' that can be set by the user. The E-step can be performed in closed form for any value of a, but the M-step is simple only if $a = 0$ or $a = 1$. The usual parameterization (3) corresponds to $a = 0$.

The parameter expansion EM (PXEM) algorithm (Liu *et al.*, 1996) uses the more general parameterization

$$y \sim N(X\beta + \theta W\gamma, \Sigma), \qquad \gamma \sim N(0, \phi^2 I),$$

where θ and ϕ are scalars and, in the notation of model (3), $\alpha = \theta\gamma$ and $\tau = \theta\phi$. In the PXEM algorithm, the vector γ is treated as missing data and the M-step maximizes over β, θ and ϕ.

We compared the old ($a = 0$), new ($a = 1$) and PXEM algorithms for a hierarchical logistic regression problem from forestry with 379 data points, three fixed effects and 15 random effects. (In this case, the algorithms are approximate EM using, at each step, the local linear approximation to the generalized linear model based on the current parameter estimates.) Fig. 7 shows the number of iterations required until convergence (defined as when the relative error for τ is less than 10^{-4}) for each algorithm, as a function of the starting value for τ. The old algorithm has the well-known problem if τ is started too low. The new algorithm corrects this, and the PXEM algorithm dominates both.

These ideas can easily be extended to multiple variance components and to the Gibbs sampler. It would also be interesting to relate to the work of Gelfand *et al.* (1994).

Peter J. Green (University of Bristol): The suggestion in the epilogue that working parameters might prove useful in Markov chain Monte Carlo (MCMC) methods is interesting, for the idea seems already much more developed there than for the EM algorithm!

Close in spirit to the formulation in Section 2, working parameters arise in auxiliary variables methods, originating in collective mode algorithms in statistical physics. In the Bayesian setting, these have potential for improving convergence and efficiency with multimodal posteriors. The idea is to enlarge the parameter space, θ becoming (θ, u), without changing the marginal posterior $p(\theta|y)$ for data y, and then to run the MCMC algorithm on (θ, u).

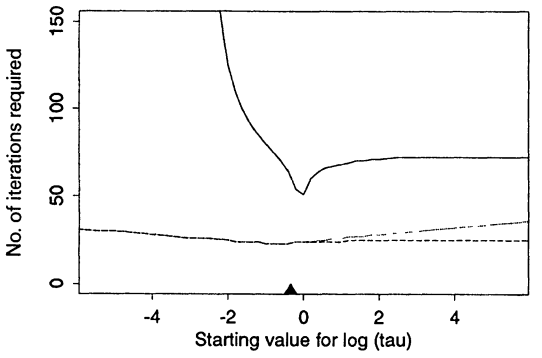

Fig. 7. Number of iterations until convergence of τ, as a function of the starting point, in the approximate EM algorithm for a hierarchical logistic problem (———, old ($a = 0$) algorithm; ·····, new ($a = 1$) algorithm; - - - -, PXEM algorithm; ▲, value of $\log \tau$ at convergence): even when τ is started at the right value, the time to convergence is non-zero because the starting values for β are not perfect

As formulated by Besag and Green (1993), the distribution of interest is

$$p(\theta) \propto \pi_0(\theta) \prod_k b_k(\theta)$$

where π_0 is of simple structure, and the 'interaction' terms $b_k(\cdot)$ are the source of slow mixing in the MCMC algorithm. We introduce auxiliary variables $\{u_k\}$, independently uniform on $[0, b_k(\theta)]$ conditional on θ. Simple manipulation reveals that

$$p(\theta|u) \propto \pi_0(\theta) \qquad \text{subject to constraints } b_k(\theta) > u_k \quad \forall k.$$

The Swendsen–Wang algorithm for the Potts and Ising models can be expressed in this way. For the *posterior* derived from the prior $\pi_0(\theta) \Pi_k b_k(\theta)$ and a likelihood $l(\theta|y)$, the idea goes through algebraically, but performance is typically severely degraded, owing to asymmetries induced by the likelihood (e.g. Gray (1993)).

By introducing a working parameter, the interactions are diluted rather than eliminated. With prior and likelihood as above, take u_k uniform on $[0, b_k(\theta)^a]$, for $a \in [0, 1]$. Then

$$p(\theta|u, y) \propto \pi_0(\theta) \prod_k b_k(\theta)^{(1-a)} l(\theta|y) \qquad \text{subject to } b_k(\theta)^a > u_k \quad \forall k:$$

compare equation (2.5).

For a toy example, consider $\theta_i \in \{-1, 0, 1\}$, $i = 1, 2$, with a Potts prior:

$$p\{\theta = (r, r)\} = \kappa\, p\{\theta = (r, s)\} \qquad \forall r \neq s,$$

following the above with $b(\theta) = \kappa^{I[\theta_1 = \theta_2]}$. Independent $y_i \in \{-1, 1\}$, $i = 1, 2$, are observed, with $l(\theta_i|y_i) \propto \lambda^{\theta_i y_i}$. There are only nine states, so the transition matrix determined by alternating Gibbs updates for u and a randomly chosen cluster can be constructed and convergence times calculated exactly. For moderately strong prior interaction ($\kappa = 4$) and uninformative data ($\lambda = 1$), the optimal a (exactly 0.5) yields convergence times 71% less than the ordinary Gibbs sampler ($a = 0$), but only 29% less than the Swendsen–Wang algorithm ($a = 1$). With stronger information in the likelihood distinguishing the parameters ($y_1 \neq y_2$, $\lambda = 2$), the optimal a is 0.3647, and savings over Gibbs and Swendsen–Wang sampling become 57% and 48%.

This picture holds on a bigger scale: Higdon (1993) reported a dramatic improvement in MCMC performance in a similar model, based on a spatial archaeology application, with 256 variables θ and a normal likelihood. (He actually used spatially inhomogeneous working parameters $a_{ij} = (1 + |y_i - y_j|)^{-1}$.) Mean times between mode swaps were around 100, compared with 1000 for the Gibbs sampler ($a = 0$), and 700 for the Swendsen–Wang algorithm ($a = 1$).

Alfred O. Hero III (University of Michigan, Ann Arbor): In this enjoyable and cogent paper the authors describe a general framework for designing fast EM algorithms which consolidates the many subset updating versions of the EM algorithm under a single umbrella. The authors' use of a 'working parameter', called a 'design parameter' in Fessler and Hero (1994a, 1995), is a key concept which we effectively employed in our development of the space alternating generalized EM (SAGE) algorithm for Poisson image reconstruction. However, the best manner to introduce this parameter into the complete–incomplete data model is strongly application dependent and it is difficult to foresee the development of any general principles for doing so. Having worked with Poisson models for over a decade, the motivation for reducing the complete data intensity via subtraction of a design parameter β_k was clear to me since it seemed the simplest way to reduce the complete data Fisher information without complicating the E-step of the SAGE algorithm. However, at least to me, the author's propitious choice of instrument $|\Sigma|^{-a}$ for introducing the working parameter a into the t-model did not seem an obvious choice.

It is remarkable that minimizing the per iteration observed information, which is related to the asymptotic convergence rate (matrix spectral radius), produced acceleration in the algorithms in both early and late iterations for these examples. However, in some applications other optimization criteria may be more directly related to the observed non-asymptotic convergence rate. For example, for smooth but possibly non-convex problems Hero and Fessler (1995) determined conditions such that the EM iterates converge monotonically in norm and, when the conditions are met, specified the contraction factor which determines the convergence rate (this analysis was applied to the SAGE algorithm in Fessler and Hero (1994b) under an interior fixed point assumption). For the Poisson model we established that a weighted Euclidean norm of the log-ratio of the iterate and the fixed point converges monotonically to 0 where the contraction factor was given by the spectral norm of a certain matrix. It is unlikely that optimizing the spectral norm and optimizing the matrix spectral radius would lead to identical values of the working parameter.

In closing, I share the authors' enthusiasm for accelerating the EM algorithm through alternating expectation–conditional maximization techniques. Alas, I suspect that the authors hopeful statement that this will require 'little additional human effort' may only apply to a limited set of clever individuals who are intimately familiar with their statistical model.

Marc Lavielle (University of Paris V): The paper summarizes various extensions of the EM algorithm to speed up its convergence. Nevertheless, these deterministic versions present other limitations:

(a) the methods require the computation of $Q(\theta|\theta^{(t)}) = E\{L(\theta|Y_{\text{aug}})|Y_{\text{obs}}; \theta^{(t)}\}$ in a closed form;
(b) convergence depends severely on the initial guess when the likelihood is not concave.

In contrast, some stochastic versions of these procedures overcome these limitations.

When the E-step is intractable, the stochastic approximation EM (SAEM) algorithm of Lavielle and Moulines (1997) splits it into a simulation step and an integration step. At iteration t, a realization $Y_{\text{aug}}^{(t)}$ of the augmented data is generated under $k(Y_{\text{aug}}|Y_{\text{obs}}; \theta^{(t)})$, the density of the augmented data given the observations and the current fit $\theta^{(t)}$. Then,

$$Q^{(t)}(\theta) = Q^{(t-1)}(\theta) + \gamma_t\{L(\theta|Y_{\text{aug}}^{(t)}) - Q^{(t-1)}(\theta)\} \tag{4}$$

is computed where $\{\gamma_t\}_{t\geqslant 0}$ is a positive and decreasing sequence. The M-step remains unchanged.

All the acceleration methods devised from the EM paradigm (see Louis (1982) or Meilijson (1989)) can be adapted to the SAEM algorithm. Of course, the E-step of the alternating expectation–conditional maximization algorithm can also be replaced by this stochastic approximation.

The main interest in the SAEM algorithm is that all the general results obtained for stochastic approximation can be used. In particular, the almost sure convergence of the sequence $\{\theta^{(t)}\}_{t\geqslant 0}$ to a stationary point of the observed data likelihood is established for a general class of complete data models. It is further demonstrated that *only* the local maxima of the incomplete likelihood are attractive for the stochastic approximation algorithm, i.e. convergence to saddlepoints is avoided with probability 1.

Furthermore, all the techniques developed to speed up convergence of stochastic algorithms can be used, such as the averaging procedure of Polyak and Juditsky (1991) and Kesten's (1958) procedure for computing an optimal sequence $\{\gamma_t\}$. They are very efficient for reducing the number of iterations of the SAEM algorithm.

In the same vein, the stochastic procedures proposed for global optimization in $\mathbb{R}^d$ can be directly adapted to the SAEM algorithm. In fact, whereas convergence of the deterministic versions of the EM

algorithm depends crucially on the initial guess $\theta^{(0)}$, a simulated annealing version of the SAEM algorithm improves convergence to the global maximum of the likelihood function. This procedure consists in adding a sequence of decreasing random variables in equation (4):

$$Q^{(t)}(\theta) = Q^{(t-1)}(\theta) + \gamma_t\{L(\theta|Y_{\text{aug}}^{(t)}) - Q^{(t-1)}(\theta)\} + c_t\eta^{(t)}. \tag{5}$$

Using the convergence results obtained by Gelfand and Mitter (1993) in a general framework, we can choose some sequences $\{\gamma_t\}$ and $\{c_t\}$ such that $\{\theta^{(t)}\}$ avoids becoming trapped at a local maximum.

Chuanhai Liu (Bell Laboratories, Murray Hill): My comments are confined to the terms *working parameter*, *simplicity* and *optimization*.

The working parameter reflects the authors' hidden simplicity conditions for both utilizing the simplicity of the EM algorithm and simplifying the brute force optimization by minimizing the fractional missing data information. In principle, a family of data augmentation (DA) schemes can be constructed and the optimal DA scheme for the fastest converging EM implementation is to be sought from the family.

Specifically, consider the following normal model with patterned covariance matrices, known as the linear mixed effects models (Laird and Ware, 1982):

$$Y_i|\theta \sim N_{n_i}(X_i\alpha,\ Z_i\Psi Z_i' + \sigma^2 I_{n_i}), \qquad \text{for } i = 1, \ldots, m,$$

where $\theta = (\alpha_{(p\times 1)},\ \Psi_{(q\times q)},\ \sigma^2_{(1\times 1)})$. By saying that a DA scheme satisfies the *simplicity conditions* for implementations of the alternating expectation–conditional maximization algorithm, I mean that the DA scheme allows for explicit closed form expressions of the E- and CM-steps. For example, the null DA scheme satisfies the simplicity conditions for the CM-step updating α with fixed (Ψ, σ^2). For easy *E-steps*, the normality in the model suggests the following general class of DA schemes:

$$Y_i^{(\text{aug})} \equiv \begin{pmatrix} Y_i \\ b_i \end{pmatrix}\Bigg|\theta \sim N_{n_i+\dim(b_i)}\left\{\begin{pmatrix} X_i\alpha \\ \mu_i(\theta) \end{pmatrix}, \begin{pmatrix} Z_i\Psi Z_i' + \sigma^2 I_{n_i} & C_i(\theta) \\ C_i'(\theta) & D_i(\theta) \end{pmatrix}\right\}, \qquad \text{for } i = 1, \ldots, m, \tag{6}$$

where $\mu_i(\theta)$, $C_i(\theta)$, $D_i(\theta)$ and $\dim(b_i)$, the dimension of the missing data b_i, are (matrix) functions of θ. The simplicity conditions for the CM-steps updating σ^2 and Ψ by maximizing the complete data log-likelihood function would require the constraints

$$\left.\begin{aligned} \dim(b_i) &= q, \\ D_i(\theta) &= H_i'(\theta)\Psi H_i(\theta), \\ C_i(\theta) &= Z_i\Psi H_i(\theta), \end{aligned}\right\} \tag{7}$$

where $H_i(\theta)$ is a $q \times q$ matrix function of θ and the corresponding CM-steps are conditional maximization given α, $\mu_i(\theta)$ and $H_i(\theta)$. For clarity, I consider fixed α and assume $Z_i'Z_i = Z_j'Z_j$ for all i, $j = 1, \ldots, m$ so that I can add the exchangeability constraints

$$\mu_i(\theta) = \mu(\theta) = 0, \qquad H_i(\theta) = H(\theta) \qquad \text{for } i = 1, \ldots, m \tag{8}$$

without worrying about losing the optimal DA scheme. Hence the family of DA schemes is defined by equations (6)–(8) or equivalently (by replacing $D_i(\theta)$ and $H_i(\theta)^{-1}$ with Ψ and $G(\theta)$ respectively) the family

$$\{(Y_i, b_i)\colon Y_i|\{b_i, \theta\} \sim N_{n_i}\{X_i\alpha + Z_iG(\theta)b_i,\ \sigma^2 I_{n_i}\} \text{ and } b_i|\theta \sim N_q(0, \Psi) \text{ for } i = 1, \ldots, m\}, \tag{9}$$

given that $G(\theta)$ satisfies the simplicity conditions for EM implementations.

Without further undesirable constraints on $G(\theta)$, it is difficult to find the optimal DA scheme analytically in family (9) by minimizing the fractional missing data information over the matrix functionals $G(\theta)$. Noticing that the purpose of the optimization is to have iterations of EM increase the actual likelihood as greatly as possible, we can view $G = G(\theta)$ as *expanded parameters* and maximize straightforwardly either the constrained actual likelihood function, as proposed by Liu (1995) for the t-model, or the constrained expected complete data likelihood, as with the parameter expansion EM algorithm (Liu *et al.*,

1996), over the expanded parameters given the current estimate $\hat{\theta}$, followed by the adjustments to $\hat{\theta}$ and the estimate $\hat{G}$ based on $\hat{G}$ with respect to the actual likelihood in such a way that $\hat{G}$ is the identity matrix for the next iteration. This approach promises *optimal* EM implementations for random effects models.

Jun S. Liu (Stanford University): Professor Meng and Professor van Dyk are to be congratulated for their eloquent paper on both historical accounts and new developments of the EM algorithm.

I concur with the authors that their ideas for the algorithm may shed light on designing data augmentation–Gibbs sampling algorithms. Whereas the convergence rate for an EM algorithm is determined by the 'fraction of missing information' (Meng and Rubin, 1994), the convergence rate for the corresponding data augmentation algorithm is governed by the 'maximal fraction of missing information', defined as the maximal correlation between y_{aug} and θ (Liu, 1994). The clever data augmenting method of Section 2 is similar to the method of parameter transformation in a Gibbs sampler for speeding up convergence. Consider designing a Gibbs sampling–data augmentation scheme by using the authors' method in Section 2. To show that the new scheme converges faster than the standard scheme, we need to verify that $y_{\text{aug}}(a)$ and θ have smaller maximal correlation than that between y_{aug} and θ. This is usually a more demanding task than just dealing with information matrices locally at the mode. In this regard, Peskun's (1973) theorem can be very powerful in giving good results (Liu, 1996). I wonder whether the authors' information matrix comparison method can be extended to comparing different Markov chain Monte Carlo (MCMC) schemes, at least approximately.

The alternating expectation–conditional maximization (AECM) algorithm reminds me of hybrid MCMC algorithms (Besag *et al.*, 1995; Green, 1995); among these, simulated tempering (Geyer and Thompson, 1995), the Swendsen–Wang algorithm and the multigrid Monte Carlo method (Goodman and Sokal, 1989) are all specially designed to take advantage of 'auxiliary' distributions or variables and to hybridize different MCMC steps. I would be interested in any comment from the authors concerning the connections between these MCMC algorithms and the AECM algorithm.

Gareth O. Roberts and Sujit K. Sahu (University of Cambridge): It is clear that the EM and Gibbs algorithms will continue to have a symbiotic relationship and the results of this paper will certainly have implications for the optimal choice of auxiliary variables in Markov chain Monte Carlo (MCMC) methods. We comment on ways in which Gibbs technology can be translated to the EM setting.

In Table 5 it is pointed out that versions of the EM algorithms have stochastic counterparts which are forms of the Gibbs sampler. When the full log-likelihood is a quadratic form in θ and $\mathbf{q}$ (for Bayesians, θ and $\mathbf{q}$ are jointly multivariate normal), the rate of convergence of the Gibbs sampler is exactly that of the partitioned EM algorithm, essentially because the conditional means and modes coincide, and the rate of convergence of Gibbs samplers on Gaussian densities is described by convergence of linear functionals (see for example Roberts and Sahu (1997)). This also allows an alternative method for calculating the rate of convergence in (for example) theorem 4.

If the target density is Gaussian with inverse dispersion matrix Q, the rate of convergence of the Gibbs sampler (and also the partitioned ECM algorithm) is the maximum modulus eigenvalue of

$$B = Q_{\text{L}}^{-1} Q_{\text{U}}$$

where Q_{L} is lower triangular and $Q = Q_{\text{L}} - Q_{\text{U}}$. A similar expression is obtained for random scan samplers. These explicit expressions allow comparisons of different blocking strategies, updating schemes and parameterizations for the Gibbs sampler. Approximation results also exist to analyse target densities which are approximately Gaussian (Roberts and Sahu, 1996).

The basic EM algorithm corresponds to the two-block Gibbs samplers. In this case, random selection of the E- or M-step is clearly never correct, but for the partitioned ECM algorithm the situation is more complicated (see Roberts and Sahu (1996, 1997)). The other versions of the Gibbs sampler cited in Table 5 are all just Gibbs samplers with different blocking and updating schemes. These can all be treated under the general framework of Roberts and Sahu (1997) and random co-ordinate selection for the conditional maximizations should often be preferred to the cyclic co-ordinate selection described by the authors.

The authors demonstrate how to improve convergence properties of the algorithms by reparameterizations. Other natural reparameterizations exist in large classes of statistical models, e.g. the stochastic *hierarchical centring* parameterizations of Gelfand *et al.* (1995).

Lastly, the authors' data augmentation technique, in particular for the t-example, induces a log-concave surface. Log-concavity of target densities has played an important role in both the theory and the practical implementaion of Gibbs sampling (see for example Polson (1996), Roberts and Tweedie (1996) and Brooks (1996)). For the EM algorithm family, log-concavity ensures convergence, but to what extent does it ensure that the convergence rates of the EM and Gibbs algorithms are approximately related? Without log-concavity, the relationship between the algorithms is less clear.

Ben Torsney (University of Glasgow): I was disappointed not to be present at the reading of the authors' paper, in contrast with 20 years ago when my contribution to Dempster *et al.* (1977) described a multiplicative algorithm for determining optimal design weights, which in the interim has been generalized.

For maximizing a general criterion $\phi(p_1, p_2, \ldots, p_J)$ subject to $p_j \geqslant 0$ and $(p_1 + p_2 + \ldots + p_J) = 1$, it takes the form

$$p_j^{(r+1)} \propto p_j^{(r)} f(d_j^{(r)}),$$

where $d_j = \partial\phi/\partial p_j$ and $f(\cdot)$ is a positive increasing function. See Silvey *et al.* (1978), Torsney (1983, 1986, 1988, 1991, 1992, 1993) and Torsney and Alahmadi (1992, 1995).

Design criteria have positive derivatives. So also does a likelihood for data from a mixture (with respect to the mixing weights), a problem which figured largely in Dempster *et al.* (1977). In its original form we had $f(d) = d^\delta$ for design criteria, where δ is a free positive parameter. This enjoys various properties. In particular the case $f(d) = d$ is monotonic for the D-optimal design criterion. It is also monotonic when the criterion is a mixture likelihood. A proof, in both cases, shows that the resultant iterations are EM iterations (see Titterington (1976) for D-optimality). What further implications for such problems (in particular the mixture problem) might be derived from the work of this further WEE GEM (see Torsney (1977)!) of a paper?

Alan M. Zaslavsky (Harvard Medical School, Boston): Listening to Meng and van Dyk's carol, I imagined the refrain, 'Why do maximum likelihood estimation at all?'. This question has special force when reading work of authors who are evidently familiar with Bayesian methods and do not have philosophical objections to prior distributions.

In their epilogue, the authors indicate two reasons why Bayesians would be interested in maximum likelihood estimates (or posterior modes, which are almost equivalent from the standpoint of computation). First, modes are useful starting values for a Markov chain Monte Carlo (MCMC) sampler (Gelman and Rubin, 1992), especially when there are multiple modes to be mapped out. Indeed, a good way to describe a multimodal posterior distribution is first to identify the modes, then to describe the local behaviour of the distribution (e.g. posterior means and covariances) around those modes and finally to compare the probability content of neighbourhoods of the different modes; a single posterior mean is a poor summary of such a distribution. The stability of the EM algorithm makes it useful for locating modes because the domains of attraction of the modes are likely to be fairly compact, whereas the behaviour of a Newton–Raphson algorithm can be chaotic when it is far from convergence.

Second, the rate of convergence of the EM algorithm is relatively easy to calculate from the sequence of iterates, whereas this is not the case for a stochastic sampler. Liu (1994) has pointed out that there are relationships between the rate of convergence of the Gibbs sampler and the corresponding EM algorithm, which may be of practical use in diagnosing convergence.

The following two situations exhibit especially close connections between an EM implementation and an MCMC sampler. When the expected complete data log-likelihood cannot be written in closed form (i.e. when there are no sufficient statistics), we may instead use a Monte Carlo EM algorithm, drawing a sample of 'missing data' and maximizing the resulting likelihood. However, having selected a value of a low dimensional parameter vector to start an MCMC routine, we should draw the starting values of the remainder of the parameters from their conditional distribution given those values. In both cases, the MCMC code can be used without modification, except that the steps involving the parameters being maximized over or fixed are replaced.

The **authors** replied later, in writing, as follows.

Besides a few expected accompaniments, the tone of the discussants is quite harmonious, especially in contrast with the (perceived) tradition of the Society. Yet, a full score of voices have been heard, from

those discussing historical trivia to those presenting new curiosities, from those verifying conjectures to those raising challenges and, of course, from those who prefer to explore a full posterior to those who just want to reach the top of one. We are grateful to the Research Section for providing us with this exceptional opportunity, as well as for inviting four lead discussants (Rubin, Titterington, Gilks and Diebolt).

We are also grateful to all the discussants for their stimulating comments, and particularly to those who have addressed the comments from other discussants. For example, Draper's question on the place of the EM algorithm in a Markov chain Monte Carlo (MCMC) world is answered by both Rubin and Zaslavsky, who also addressed the issue of multimodality of a likelihood raised by Diebolt, Chauveau, Dupuis and Lavielle. It has been repeatedly emphasized in the literature (e.g. Meng and Rubin (1992)) that, in the presence of several (important) modes, avoiding local modes, while requiring numerical sophistication, is statistically naïve. Smith's examples provide a perfect illustration of this point. Somewhat ironically, the use of MCMC methods should help to make this point more transparent than ever.

As another example, Hinde's search for the optimal working parameter a in the mixed effect model was essentially carried out by Aitkin, whose result independently verifies part of the results that we obtained in Meng and van Dyk (1997). We anticipated that the mixed effect model application would be of great practical interest but had no space for it. We are thus particularly pleased to see that this application received comments from several discussants (Aitkin, Gelman, Hinde, C. Liu and Rubin). Space limitations also led to deletions of important historical references, as Rubin noted. A more detailed account of the earliest known (at least to us) history of the EM algorithm is now available in Meng (1997a).

It is evident that the two key messages of our paper, namely

(a) EM-type algorithms can be made much faster with little cost in terms of simplicity or stability and
(b) what is useful for speeding up EM-type algorithms may be suggestive to MCMC methods,

have been well received. Below we would like to discuss several issues in more detail than we could in the paper and at the same time examine some stimulating points from the discussants. We attempt to provide something for everybody, from the casual reader to the specialist. As a consequence, an average reader might find parts of our discussion too detailed and parts too dense. We also try to follow the tradition of the Society — to be critical to stimulate the exchange of ideas. We list a discussant's name in parentheses whenever a point that we discuss is related to the discussant's comments, and each discussant is mentioned at least once in our reply.

Interplay of EM and Markov chain Monte Carlo: efficient data augmentation and reparameterization

The interplay between EM and MCMC is a topic discussed by many and is summarized well by Gilks's first-rate contribution. (The most an author can hope from a discussant is a full day of programming to verify the author's conjecture!) The ideas of using auxiliary variables and efficient (re)parameterizations are almost as old as the MCMC method itself — data augmentation is a form of auxiliary variables (Gilks, Green, J. Liu and Roberts and Sahu). What we emphasized is that it might be useful

(a) to look for 'unnatural or unfamiliar' parameterizations and
(b) to consider first the rate of convergence of EM-type algorithms even if we are only interested in accelerating MCMC convergence.

What then are the general rules for seeking unnatural or unfamiliar data augmentation or parameterizations (Gilks, Hero and Kent)? The phrase unnatural or unfamiliar suggests that such a search must be a matter of art or creativity, at least for now. Nevertheless, there are common themes underlying the successful applications that we are aware of. Both the t-application and the mixed effect application are examples of *rescaling*, and the Poisson application as well as the parameterization reported in Gelfand *et al.* (1995) are cases of *recentring*. Thus these two schemes seem to be a good starting point, with possible extension from scaling to scaling and rotation in multivariate cases. Even the use of a working parameter in the exponent is not completely 'out of the blue' (though we did 'dream it up'), as Torsney's contribution seems to suggest.

In addition, C. Liu's approach seems quite promising. Liu illustrated that the idea of restricting the data augmentation scheme by the *simplicity* requirement can be formulated mathematically to reduce a natural looking but overly large family. One then maximizes over any 'left-over' free structure (i.e. the working parameter) *within* the iterations. As summarized by Rubin, this maximization approach complements our

approach which seeks an optimal value of the working parameter analytically *outside* the iteration. The advantage of Liu's approach, or more generally the expanded parameter approach (Rubin), is its flexibility in implementing the (adaptive) optimal algorithm as it effectively estimates the optimal expanded parameter numerically within the algorithm. This overcomes some difficulties that we have encountered. For example, it allows an EM-type implementation for the mixed effects model which works well relative to the standard algorithm for any parameter value, in contrast with our algorithm which is optimal with a restricted class of data augmentation schemes (i.e. $a = 0$ or $a = 1$ in Hinde's notation; see Meng and van Dyk (1997)) but depends on the parameter value. Preliminary theory (van Dyk and Meng, 1997) shows that the parameter expansion and efficient data augmentation approaches will produce the same optimal theoretical rate of convergence, assuming a suitable correspondence between the underlying data augmentations or parameterizations, confirming Rubin's insight. In other words, the differences given in Rubin's Table 6 can be largely explained by a table similar to Aitkin's Table 8, using the theoretical approximation that the number of iterations is proportional to the negative of the reciprocal of the logarithm of the rate of the convergence (i.e. the fraction of missing information).

We note that C. Liu's motivation for maximizing over the working parameter along the iteration came from the intuition that the more we increase the likelihood the faster the resulting algorithm. This is true when we directly increase the actual likelihood, but with EM-type algorithms it is inevitable that we shall only be able to increase the actual likelihood indirectly, at least in some of the CM-steps, by increasing the Q-function defined by equation (3.1). Interestingly, as we alluded to in Section 3.1, with this indirect approach, Meng (1994) has shown that the intuition can be misleading—an expectation–conditional maximization (ECM) algorithm can be faster than the corresponding EM algorithm even though the latter always increases Q more at each iteration. Furthermore, more frequent updating of the E-step within each iteration (i.e. multicycle ECM) does not necessarily lead to a better rate of convergence, again contrary to common wisdom. These issues also appear, not surprisingly, in the Gibbs sampler context; see, for example, Liu *et al.* (1994). As far as we know, these counter-intuitive phenomena have not been well understood within the EM framework and we certainly hope that the study of MCMC schemes, particularly the sampling schemes involving the Gibbs sampler, can shed some light (see the warning listed below, however) (Roberts and Sahu).

Understanding differences and limitations

An effective utilization of the interplay between EM and MCMC methods requires an understanding of the differences between them beyond the obvious ones. Gilks's t-model implementation is a good illustration. An effective data augmentation scheme for EM algorithms (e.g. a closed form mode) may not be effective for MCMC algorithms (e.g. an easy sampling scheme), and vice versa. Furthermore, Gilks's comment on balancing computational saving with increased human effort is right on target, a point that we also emphasized in Section 3.1 and elsewhere (e.g. Meng and Schilling (1996), Meng (1997a, b) and van Dyk and Meng (1997)). Of course, we should distinguish the human effort of those who design algorithms from that of general users. Sometimes it pays (and is necessary) to increase the former to decrease the latter. For example, once the code is available, the Gibbs sampler using the new parameterization of the t-model can be advantageous for a user who chooses to stop the iteration at a prespecified number of iterations, which although a questionable procedure is quite common in practice. (We are pleasantly surprised that the new Gibbs sampler did not take more central processor unit time per iteration—apparently this is due to the effectiveness of Gilks's adaptive rejection Metropolis sampling method.) The human effort discussed in our paper refers to the latter, the effort required by general users, as emphasized in Section 3.5 by contrasting the simplicity of the resulting Fessler–Hero code with the lengthy derivations behind them (Hero).

A second difference between EM and MCMC algorithms is that directly imputing the missing or latent variable with one random draw, as in stochastic EM, does not work (well) within the EM framework in the sense that the resulting algorithm does not produce a (local or global) maximum likelihood estimator (MLE) or a sequence of draws from a desired posterior (Gilks). In other words, not all cross-fertilizations are equally fruitful. Although various *ad hoc* approaches can be useful for initial exploration, it is important to understand their limitations, especially those that may have adverse effects (e.g. missing important local modes). Indeed, the same warning applies to less *ad hoc* methods as well, such as EM itself. For example, our experience suggests that it is often not productive, at least not as productive, to apply the EM algorithm to cases where it loses its principal advantages, such as when the augmented data density is not from an exponential family (Chauveau). Unproductive or misguided use seems to be an unavoidable side-effect of a popular method, be it statistical or computational. To reduce such misuses, those of us who have invested heavily in a particular methodology should be

among the first to provide explicit warnings regarding the limitations of the methodology. The entry on the EM algorithm in the *Encyclopedia of Statistical Sciences* (Meng, 1997b) carries such warnings. A truly general approach is indeed only a theoretical possibility (Damien and Walker).

On a closely related point, a casual reader may misread Gilks's comparison of the EM algorithm and the Gibbs sampler as saying that the former directly replaces the missing data by their conditional expectation. We know that this generally will not lead to the MLE or even a consistent estimator—the key here is that Gilks assumes that $p(Y_{\text{aug}}|\theta)$ is from an exponential family and Y_{aug} is *sufficient* for θ. Strictly, the condition *sufficient* should at least be replaced with *minimally sufficient* since the data themselves are always sufficient. But even saying that EM replaces the minimum sufficient statistics by their conditional expectations is still not correct because minimal sufficiency is invariant under one-to-one mapping of the statistics. The correct condition for legitimate direct imputation of the missing or latent data with their conditional expectation is that $\log p(Y_{\text{aug}}|\theta)$ is linear in the *unobserved part* of Y_{aug}. Because directly imputing the missing data is intuitively so appealing (indeed, it was the predecessor of the EM algorithm; see Meng and Rubin (1992)), on too many occasions we had to explain that directly imputing the missing data will generally not yield a local or a global MLE. For this reason, we take this opportunity to clarify this distinction further, and to make sure that it is generally understood that the traditional use of the term 'sufficient' in the EM literature regarding this issue, as reflected in Gilks's as well as our own writings, really means the aforementioned linearity.

A third (potential) difference is with respect to efficient orderings for executing the conditional maximizations (e.g. for ECM) and of conditional sampling schemes (for the Gibbs sampler) (Roberts and Sahu). Initially, we had the same intuition as Roberts and Sahu, based on the work by Amit and Grenander (1991) and the similarity between the rate of convergence of a partitioned ECM algorithm and the rate of convergence of a Gibbs sampler (see Meng (1990, 1994)). (Incidentally, the expression given by Roberts and Sahu is for the rate of convergence of the partitioned CM algorithm, which has a direct symbolic analogy with the Gibbs sampler. The corresponding expression for convergence of the partitioned ECM algorithm, as a special case of our theorem 4, is more complicated owing to the presence of the E-step; see Meng (1990) or van Dyk *et al.* (1995) for an explicit formula using triangular matrices.) That is, we expected that the random co-ordinate selection would be more efficient than cyclic co-ordinate selection, at least on average. Our empirical studies, reported in van Dyk and Meng (1997), however, did not provide any evidence for this preference when using ECM-type algorithms. We in fact found, in our simulations, that cyclic co-ordinate selection is slightly more effective on average. In addition, we found that, although reversing the order of conditional maximizations has no effect on the theoretical rate of convergence of the ECM algorithm, such reversals can have quite a noticeable effect on the actual number of iterations required for convergence. These findings alerted us one more time that it is generally wise to conduct realistic empirical evaluations of a strategy before adopting it, even if intuition, common wisdom and theory (they are often highly dependent!) all point to one strategy. Given these somewhat unexpected findings, we are very interested in seeing Roberts and Sahu's general framework, which may provide some insight. Of course, our interest is more at the conceptual and theoretical level, as the gain in efficiency from sensible ordering is likely to be secondary compared with that from efficient data augmentation or parameterization.

Finally, since we assume essentially the same conditions as Dempster *et al.* (1977) and Wu (1983), our theorems 3 and 4 are applicable only when the convergence points are inside the parameter space, as we made explicit in Section 3.3 (Fessler). (It is also important to note that our theorem 4 does not imply that AECM is faster than EM (Diebolt).) The importance of recognizing this restriction is not only that convergence to boundary points is not merely a mathematical possibility but also that, once it happens, the usual maximum likelihood inference based on quadratic approximation is generally inadequate (Smith). This has been a recurrent theme in statistics, i.e. when there is a computational or mathematical problem there often is a more fundamental statistical problem underlying it. Interestingly, for the mixed effects model, the data augmentation with $a = 1$, i.e. Hinde's z-formulation, effectively eliminates the boundary problem because τ does not need to be positive even if there is no information in the observed data to estimate its sign (i.e. we effectively have included an expanded parameter $\xi = \text{sgn}(\tau)$; we are only interested in τ^2, the random effect variance). In Meng and van Dyk (1997), we argued that avoiding the boundary problem may explain the fast convergence of the new algorithm when the estimate of τ^2 is relatively small. Because a very small estimate of τ^2 should automatically warn the investigator of the possible misspecification of the model (see the discussion of Gelman *et al.* (1996) by B. Hill and the reply), the model diagnostic information in the speed is not lost when we move from a slower to a faster implementation, as long as the estimates are correctly calculated.

Seeking new fronts and new insight

Rubin's comment that there is still much to do is evident from the discussion, even if we exclude everything about computational efficiency. EM-type algorithms, given their success so far, will certainly find even more applications, either with specific models or in a whole new field (Titterington). On the methodological side, there have been efforts to extend the central idea of the EM algorithm beyond likelihood or Bayesian computation, for example, with general estimating equations (e.g. Heyde and Morton (1996)). The computation for M-estimates, as discussed by Tyler, is another possible area. Although we do not have an answer for Tyler's general question, we do have an observation that might be useful for going beyond the maximum likelihood framework.

Although we almost always discuss the EM algorithm assuming that $f(Y_{\text{aug}}|\theta)$ and $f(Y_{\text{obs}}|\theta)$ are the augmented data and observed data *density* or *probability* functions, a closer examination of the general theory underlying EM (e.g. Dempster *et al.* (1977) and Wu (1983)) would reveal that it is not necessary to require either of them to be a proper density or probability function. The only requirement is that $f(Y_{\text{aug}}|\theta)/f(Y_{\text{obs}}|\theta)$ is a *proper* conditional density or probability function $f(Y_{\text{aug}}|Y_{\text{obs}}, \theta)$. This fact, though trivial, is important for justifying EM in many (empirical) Bayesian calculations when improper (prior) densities are involved. It is even important for some likelihood calculations, e.g. for the restricted maximum likelihood estimates with mixed effects models (Meng and van Dyk, 1997). The point is that it is legitimate to replace $f(Y_{\text{aug}}|\theta)$ and $f(Y_{\text{obs}}|\theta)$ by general objective functions, say $D(Y_{\text{aug}}, \theta)$ and $D(Y_{\text{obs}}, \theta)$, as long as $D(Y_{\text{aug}}, \theta)/D(Y_{\text{obs}}, \theta)$ can be treated as a proper conditional density or probability function in place of $f(Y_{\text{aug}}|Y_{\text{obs}}, \theta)$.

Also viewing the E-step as a maximization could be another important step in exploring new fronts (Titterington). Minimally, this should help to make the symbolic connection between the EM algorithm and the Gibbs sampler more direct as the latter only involves one operator: sampling (Roberts and Sahu). We look forward to development in this direction, and we thank Titterington for the introduction and, of course, for his informative discussion and many new references — the task of Meng and Pedlow (1992) is now beyond impossible!

We also had the same wish as Titterington, i.e. to understand why the optimal data augmentation scheme with the t-model is one of the few that can be implemented. We are still amused that the 'free lunch' turns out to be (almost) the only 'eatable lunch'! We very much looked forward to insight from the discussants, but we must confess that we are now even more puzzled. The mysterious optimal working parameter value $1/(\nu + p)$ for the t-model now appears as the breakdown point of the MLE when $\nu \geq 1$ (Tyler) and as the breakdown point (in its literal sense!) of a log-concavity (Gilks). Are these all simply mathematical coincidences or do we simply lack imagination? Could this be a problem for Hilbert's proverbial list for the next century (or even the next millennium)?

REFERENCES IN THE DISCUSSION

Amari, S. (1995) Information geometry of the EM and em algorithms for neural networks. *Neur. Netwks*, **8**, 1379–1408.

Amit, Y. and Grenander, U. (1991) Comparing sweep strategies for stochastic relaxation. *J. Multiv. Anal.*, **37**, 197–222.

Anderson, D. A. and Hinde, J. P. (1988) Random effects in generalized linear models and the EM algorithm. *Communs Statist. Theory Meth.*, **17**, 3847–3856.

Anderson, N. H. and Titterington, D. M. (1995) Beyond the binary Boltzmann machine. *IEEE Trans. Neur. Netwks*, **6**, 1229–1236.

Archer, G. E. B. and Titterington, D. M. (1996) Parameter estimation for hidden Markov chains. To be published.

Arslan, O., Constable, P. D. L. and Kent, J. T. (1995) Convergence behaviour of the EM algorithm for the multivariate t-distribution. *Communs Statist. Theory Meth.*, **24**, 2981–3000.

Beale, E. M. L. and Little, R. J. A. (1975) Missing values in multivariate analysis. *J. R. Statist. Soc.* B, **37**, 129–145.

Besag, J. and Green, P. J. (1993) Spatial statistics and Bayesian computation. *J. R. Statist. Soc.* B, **55**, 25–37.

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statist. Sci.*, **10**, 3–66.

Biscarat, J.-C. (1994) Almost sure convergence of a class of stochastic algorithms. *Stoch. Processes Appl.*, **50**, 83–99.

Bock, R. D. and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of an EM algorithm. *Psychometrika*, **46**, 443–459.

Brooks, S. P. (1996) Convergence diagnostics for Markov chain Monte Carlo. *PhD Thesis*. Statistical Laboratory, University of Cambridge, Cambridge.

Byrne, C. (1993) Iterative image reconstruction algorithms based on cross-entropy minimization. *IEEE Trans. Image Process.*, **2**, 106–114.

———(1996) Block-iterative methods for image reconstruction from projections. *IEEE Trans. Image Process.*, **5**, 792–794.

Byrne, W. (1992) Alternating minimization and Boltzmann machine learning. *IEEE Trans. Neur. Netwks*, **3**, 612–620.

Celeux, G., Chauveau, D. and Diebolt, J. (1996) Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Statist. Comput. Simuln*, **55**, 287–314.

Celeux, G. and Diebolt, J. (1985) The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Comput. Statist. Q.*, **2**, 73–82.

———(1988) A probabilistic teacher algorithm for iterative maximum likelihood estimation. In *Classification and Related Methods of Data Analysis* (ed. H. H. Bock), pp. 617–623. Amsterdam: North-Holland.

———(1992) A stochastic approximation type EM algorithm for the mixture problem. *Stoch. Stoch. Rep.*, **41**, 127–146.

Ceppellini, R., Siniscalco, M. and Smith, C. A. B. (1955) The estimation of gene frequencies in a random mating population. *Ann. Hum. Genet.*, **20**, 97–115.

Chauveau, D. (1992) Algorithmes EM et SEM pour un mélange censuré de distributions de défaillances, application à la fiabilité. *Rev. Statist. Appl.*, **40**, 67–76.

———(1995) A stochastic EM algorithm for mixtures with censored data. *J. Statist. Planng Inf.*, **46**, 1–25.

Csiszár, I. and Tusnády, G. (1984) Information geometry and alternating minimization procedures. In *Statistics and Decisions* (eds E. J. Dudewicz *et al.*), suppl. issue 1, pp. 205–237. Munich: Oldenburg.

Cumbus, C., Damien, P. and Walker, S. G. (1996a) Sampling from nonstandard distributions via the Gibbs sampler. To be published.

Damien, P. and Walker, S. G. (1996a) Sampling probability densities via uniform random variables and a Gibbs sampler. To be published.

———(1996b) A full Bayesian analysis of circular data using the von Mises distribution. To be published.

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

Diebolt, J. and Celeux, G. (1993) Asymptotic properties of a stochastic EM algorithm for estimating mixture proportions. *Stoch. Mod.*, **9**, 599–613.

Diebolt, J. and Ip, E. H. S. (1996) A stochastic EM algorithm for approximating the maximum likelihood estimate. In *Markov Chain Monte Carlo in Practice* (eds W. R. Gilks, S. T. Richardson and D. J. Spiegelhalter). London: Chapman and Hall.

Draper, D. and Cheal, R. (1997) Causal inference via Markov Chain Monte Carlo. *Technical Report*. Statistics Group, University of Bath, Bath.

Dunmur, A. P. and Titterington, D. M. (1997) Analysis of latent structure models with multidimensional latent variables. In *Statistics and Neural Networks: Recent Advances at the Interface* (eds J. W. Kay and D. M. Titterington). Oxford: Oxford University Press. To be published.

Dupuis, J. A. (1995) Bayesian estimation of movement and survival probabilities from capture–recapture data. *Biometrika*, **82**, 761–772.

van Dyk, D. A. and Meng, X. L. (1997) The art of data augmentation. *Technical Report*. Department of Statistics, University of Chicago, Chicago.

van Dyk, D. A., Meng, X. L. and Rubin, D. B. (1995) Maximum likelihood estimation via the ECM algorithm: computing the asymptotic variance. *Statist. Sin.*, **5**, 55–75.

Fessler, J. A., Clinthorne, N. H. and Rogers, W. L. (1993) On complete data spaces for PET reconstruction algorithms. *IEEE Trans. Nucl. Sci.*, **40**, 1055–1061.

Fessler, J. A., Ficaro, E. P., Clinthorne, N. H. and Lange, K. (1997) Grouped-coordinate ascent algorithms for penalized-likelihood transmission image reconstruction. *IEEE Trans. Med. Imgng*, **16**, in the press.

Fessler, J. A. and Hero III, A. O. (1994a) Space-alternating generalized EM algorithms for penalized maximum-likelihood image reconstruction. *Technical Report 286*. Communications and Signal Processing Laboratory, University of Michigan, Ann Arbor.

———(1994b) Space alternating generalized expectation maximization algorithm. *IEEE Trans. Signal Process.*, **42**, 2664–2677.

———(1995) Penalized maximum likelihood image reconstruction using space alternating generalized EM algorithms. *IEEE Trans. Image Process.*, **4**, 1417–1429.

Fessler, J. A. and Rogers, W. L. (1996) Spatial resolution properties of penalized-likelihood image reconstruction methods: space-invariant tomographs. *IEEE Trans. Image Process.*, **5**, 1346–1358.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parametrization for normal linear mixed models. *Biometrika*, **82**, 479–488.

Gelfand, S. B. and Mitter, S. K. (1993) Metropolis-type annealing algorithms for global optimization in $\mathbb{R}^d$. *SIAM J. Control Optimizn*, **31**, 111–131.

Gelman, A., Meng, X. L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). *Statist. Sin.*, **6**, 733–807.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences (with discussion). *Statist. Sci.*, **7**, 457–472.

Geyer, C. J. and Thompson, E. A. (1995) Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Am. Statist. Ass.*, **90**, 909–920.

Ghahramani, Z. (1995) Factorial learning and the EM algorithm. In *Advances in Neural Information Processing Systems 7* (eds G. Tesauro, D. S. Touretzky and T. K. Leen). Cambridge: MIT Press.

Gilks, W. R., Best, N. G. and Tan, K. K. C. (1995) Adaptive rejection Metropolis sampling within Gibbs sampling. *Appl. Statist.*, **44**, 455–472.

Goodman, J. and Sokal, A. D. (1989) Multigrid Monte Carlo method: conceptual foundations. *Phys. Rev.* D, **40**, 2035–2071.

Gouno, E. (1996) Problèmes de fiabilité issus de l'industrie: méthodes algorithmiques, méthodes Bayésiennes. *PhD Thesis.* Université de Marne la Vallée, Noisy-le-Grand.

Gray, A. J. (1993) Discussion on The Gibbs sampler and other Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 58–61.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Guo, S. W. and Thompson, E. A. (1994) Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics*, **50**, 417–432.

Hathaway, R. J. (1986) Another interpretation of the EM algorithm for mixture distributions. *Statist. Probab. Lett.*, **4**, 53–56.

Hero III, A. O. and Fessler, J. A. (1995) Convergence in norm for alternating expectation-maximization (EM) type algorithms. *Statist. Sin.*, **5**, 41–54.

Heyde, C. C. and Morton, R. (1996) Quasi-likelihood and generalizing the EM algorithm. *J. R. Statist. Soc.* B, **58**, 317–327.

Higdon, D. (1993) Discussion on The Gibbs sampler and other Markov chain Monte Carlo methods. *J. R. Statist. Soc.* B, **55**, 78.

Hudson, H. M. and Larkin, R. S. (1994) Accelerated image reconstruction using ordered subsets of projection data. *IEEE Trans. Med. Imgng*, **13**, 601–609.

Ip, E. H. S. (1994) *PhD Thesis.* Stanford University, Stanford.

Kent, J. T. and Tyler, D. E. (1991) Redescending M-estimates of multivariate location and scatter. *Ann. Statist.*, **19**, 2102–2119.

———(1996) Constrained M-estimation for multivariate location and scatter. *Ann. Statist.*, **24**, 1346–1370.

Kent, J. T., Tyler, D. E. and Vardi, Y. (1994) A curious likelihood identity for the multivariate *t*-distribution. *Communs Statist. Simuln*, **23**, 441–453.

Kesten, H. (1958) Accelerated stochastic approximation. *Ann. Math. Statist.*, **29**, 41–59.

Laird, N. M. and Ware, J. H. (1982) Random effects models for longitudinal data. *Biometrics*, **38**, 967–974.

Lavielle, M. and Moulines, E. (1997) A simulated annealing version of the EM algorithm for non-Gaussian deconvolution. *Statist. Comput.*, to be published.

Liu, C. (1995) On maximum likelihood estimation of the *t* distribution using EM algorithms. *Technical Report.* Bell Laboratories, Murray Hill.

Liu, C., Rubin, D. B. and Wu, Y. (1996) Parameter expansion for EM acceleration: the PX-EM algorithm. Submitted to *Biometrika.*

Liu, J. S. (1994) The fraction of missing information and convergence rate for data augmentation. In *Computing Science and Statistics: Proc. 26th Symp. Interface* (eds J. Sall and A. Lehman), pp. 490–497. Research Triangle: Interface Foundation of North America.

———(1996) Peskun's theorem and a modified discrete-state Gibbs sampler. *Biometrika*, **83**, 681–682.

Liu, J. S., Wong, W. H. and Kong, A. (1994) Covariance structure of the Gibbs sampler with applications to comparisons of estimators and augmentation schemes. *Biometrika*, **81**, 27–40.

Lopuhaä, H. P. (1989) On the relationship between S-estimators and M-estimators of multivariate location and covariance. *Ann. Statist.*, **17**, 1661–1683.

Louis, T. A. (1982) Finding the observed information matrix when using the EM algorithm. *J. R. Statist. Soc.* B, **44**, 226–233.

Maronna, R. A. (1976) Robust M-estimates of multivariate location and scatter. *Ann. Statist.*, **4**, 51–67.

McCulloch, C. (1994) Maximum likelihood variance components estimation for binary data. *J. Am. Statist. Ass.*, **89**, 330–335.

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

Meilijson, I. (1989) A fast improvement to the EM algorithm on its own terms. *J. R. Statist. Soc.* B, **51**, 127–138.

Meng, X. L. (1990) Towards complete results for some incomplete data problems. *PhD Thesis.* Department of Statistics, Harvard University, Cambridge.

———(1994) On the rate of convergence of the ECM algorithm. *Ann. Statist.*, **22**, 326–339.

Meng, X. L. (1997a) The EM algorithm and medical studies: a historical link. *Statist. Meth. Med. Res.*, **6**, in the press.

———(1997b) The EM algorithm. In *Encyclopedia of Statistical Sciences* (ed. S. Kotz), suppl. vol. New York: Wiley. To be published.

Meng, X. L. and van Dyk, D. A. (1997) Fast EM-type implementations for mixed effects models. Submitted to *J. R. Statist. Soc.* B.

Meng, X. L. and Pedlow, S. (1992) EM: a bibliographic review with missing articles. *Proc. Statist. Comput. Sect. Am. Statist. Ass.*, 24–27.

Meng, X. L. and Rubin, D. B. (1992) Recent extensions to the EM algorithm (with discussion). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 307–320. Oxford: Oxford University Press.

———(1994) On the global and componentwise rates of convergence of the EM algorithm. *Lin. Alg. Applic.*, **199**, 413–425.

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

Morgan, B. J. T. and Titterington, D. M. (1977) A comparison of iterative methods for obtaining maximum-likelihood estimates in contingency tables with a missing diagonal. *Biometrika*, **64**, 265–269.

Morganthaler, S. (1990) Fitting redescending M estimators in regression. In *Robust Regression: Analysis and Applications* (eds K. D. Lawrence and J. L. Arthur), pp. 105–128. New York: Dekker.

Neal, R. M. and Hinton, G. E. (1993) A new view of the EM algorithm that justifies incremental and other variants. *Technical Report*. Department of Computer Science, University of Toronto, Toronto.

Orchard, T. and Woodbury, M. A. (1972) A missing information principle: theory and applications. In *Proc. 6th Berkeley Symp. Mathematical Statistics and Probability*, vol. 1, pp. 697–715. Berkeley: University of California Press.

Peskun, P. (1973) Optimum Monte Carlo sampling using Markov chains. *Biometrika*, **60**, 607–612.

Polson, N. G. (1996) Convergence of Markov chain Monte Carlo algorithms. In *Bayesian Statistics 5* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 297–321. Oxford: Oxford University Press.

Polyak, B. T. and Juditsky, A. B. (1991) Acceleration of stochastic approximation by averaging. *SIAM J. Control Optimizn*, **29**, 838–855.

Qian, W. and Titterington, D. M. (1992) Estimation of parameters in hidden Markov models. *Phil. Trans. R. Soc. Lond.* A, **337**, 407–428.

Redner, R. A. and Walker, H. F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.*, **26**, 195–239.

Robert, C. P. (1992) Discussion on Recent extensions to the EM algorithm (by X. L. Meng and D. B. Rubin). In *Bayesian Statistics 4* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.

Robert, C. P., Celeux, G. and Diebolt, J. (1993) Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statist. Probab. Lett.*, **16**, 77–83.

Robert, C. P. and Soubiran, C. (1993) Estimation of a mixture model through Bayesian sampling and prior feedback. *TEST*, **2**, 125–146.

Roberts, G. O. and Sahu, S. K. (1996) Rate of convergence of the Gibbs sampler by Gaussian approximation. *Technical Report 96-21*. Statistical Laboratory, University of Cambridge, Cambridge.

———(1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *J. R. Statist. Soc.* B, **59**, 291–317.

Roberts, G. O. and Tweedie, R. L. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

Rubin, D. B. (1983) Iteratively reweighted least squares. In *Encyclopedia of Statistical Sciences* (eds S. Kotz and N. L. Johnson), vol. 4, pp. 272–275. New York: Wiley.

———(1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann. Statist.*, **12**, 1151–1172.

Ruppert, D. (1992) Computing S estimators for regression and multivariate location/dispersion. *J. Comput. Graph. Statist.*, **1**, 253–370.

Schull, W. J. and Ito, Y. U. (1969) Note on the estimation of the ABO gene frequencies and the coefficient of inbreeding. *Am. J. Hum. Genet.*, **1**, 168–177.

Shepp, L. A. and Vardi, Y. (1982) Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imgng*, **1**, 113–122.

Silvey, S. D., Titterington, D. M. and Torsney, B. (1978) An algorithm for optimal design on a finite design space. *Communs Statist.*, **14**, 1379–1389.

Smith, C. A. B. (1957) Counting methods in genetical statistics. *Ann. Hum. Genet.*, **21**, 254–276.

———(1969) *Biomathematics*, vol. 2. London: Griffin.

———(1997) Obituary: Maurice Charles Kenneth Tweedie. *J. R. Statist. Soc.* A, **160**, 151–154.

Smith, C. A. B. and Stephens, D. A. (1996) Estimating linkage heterogeneity. *Ann. Hum. Genet.*, **60**, 161–169.

Sundberg, R. (1974) Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, **1**, 49–58.

Tatsuoka, K. (1996) M, CM, and S-estimates: theory and computation. *PhD Thesis*. Department of Statistics, Rutgers University, New Brunswick.

Titterington, D. M. (1976) Algorithms for computing D-optimal designs on a finite design space. In *Proc. Conf. Information Sciences and Systems*, pp. 213–216. Baltimore: Johns Hopkins University Press.

Torsney, B. (1977) Discussion on Maximum likelihood from incomplete data via the EM algorithm (by A. P. Dempster, N. M. Laird and D. B. Rubin). *J. R. Statist. Soc.* B, **39**, 26–27.

———(1983) A moment inequality and monotonicity of an algorithm. *Lect. Notes Econ. Math. Syst.*, **215**, 249–260.

———(1986) Discussion on On the statistical analysis of dirty pictures (by J. Besag). *J. R. Statist. Soc.* B, **48**, 296.

———(1988) Computing optimizing distributions with applications in design, estimation and image processing. In *Optimal Design and Analysis of Experiments* (eds Y. Dodge, V. V. Fedorov and H. P. Wynn), pp. 361–370. Amsterdam: Elsevier Science.

———(1991) Discussion on Empirical functionals and efficient smoothing parameter selection (by P. Hall and I. Johnstone). *J. R. Statist. Soc.* B, **53**, 513–514.

———(1992) Discussion on Constrained Monte Carlo maximum likelihood for dependent data (by C. J. Geyer and E. A. Thompson). *J. R. Statist. Soc.* B, **54**, 688–689.

———(1993) Discussion on From image deblurring to optimal investments: maximum likelihood solutions for positive linear inverse problems (by Y. Vardi and D. Lee). *J. R. Statist. Soc.* B, **55**, 601–602.

Torsney, B. and Alahmadi, A. M. (1992) Further development of algorithms for constructing optimizing distributions. In *Model Oriented Data Analysis: Proc. 2nd IIASA Wkshp, St Kyrik* (eds V. V. Federov, W. G. Müller and I. N. Vuchkov), pp. 121–129. Vienna: Physica.

———(1995) Designing for minimally dependent observations. *Statist. Sin.*, **5**, 405–419.

Walker, S. G. and Damien, P. (1996a) Sampling a Dirichlet process mixture model. To be published.

———(1996b) A full Bayesian analysis of a neutral to the right process. To be published.

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

Woodruff, D. L. and Rocke, D. M. (1994) Computable robust estimation of multivariate location and shape in high dimensions using compound estimators. *J. Am. Statist. Ass.*, **89**, 888–896.

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

Xu, L. and Jordan, M. I. (1996) On convergence properties of the EM algorithm for Gaussian mixtures. *Neur. Comput.*, **8**, 129–151.

Zhang, J. (1992) The Mean Field Theory in EM procedures for Markov random fields. *IEEE Trans. Signal Process.*, **40**, 2570–2583.

———(1993) The Mean Field Theory in EM procedures for blind Markov random field image restoration. *IEEE Trans. Image Process.*, **2**, 27–40.