Scandinavian Journal of Statistics, Vol. 35: 335–353, 2008
doi: 10.1111/j.1467-9469.2007.00585.x

Simple and Globally Convergent Methods for Accelerating the Convergence of Any EM Algorithm

RAVI VARADHAN

The Center on Aging and Health, Johns Hopkins University

CHRISTOPHE ROLAND

Laboratoire Paul Painlevé, Université des Sciences et Technologies de Lille

ABSTRACT. The expectation-maximization (EM) algorithm is a popular approach for obtaining maximum likelihood estimates in incomplete data problems because of its simplicity and stability (e.g. monotonic increase of likelihood). However, in many applications the stability of EM is attained at the expense of slow, linear convergence. We have developed a new class of iterative schemes, called squared iterative methods (SQUAREM), to accelerate EM, without compromising on simplicity and stability. SQUAREM generally achieves superlinear convergence in problems with a large fraction of missing information. Globally convergent schemes are easily obtained by viewing SQUAREM as a continuation of EM. SQUAREM is especially attractive in high-dimensional problems, and in problems where model-specific analytic insights are not available. SQUAREM can be readily implemented as an 'off-the-shelf' accelerator of any EM-type algorithm, as it only requires the EM parameter updating. We present four examples to demonstrate the effectiveness of SQUAREM. A general-purpose implementation (written in R) is available.

Key words: causal inference, conjugate gradient, EM acceleration, finite mixtures, fixed point iteration, quasi-Newton, squared iterative method

1. Motivation

The expectation-maximization (EM) algorithm (Dempster *et al.*, 1977) is a ubiquitous approach for obtaining maximum likelihood estimates (MLEs) in incomplete data problems occurring in a wide array of applications. The key notion in the use of EM is that while the direct maximization of the likelihood for the observed data is difficult, augmenting the observed data with the missing information results in a simpler maximization problem. The popularity of EM is due to its simplicity and stability (monotonic increase of likelihood). However, in many applications the stability of EM is achieved at the expense of slow, linear convergence, where the rate of convergence is inversely related to the proportion of missing information. Such slow convergence limits the ability of the analyst to perform computer-intensive tasks such as model exploration and validation, sensitivity analysis, bootstrapping and simulations. Slow convergence also limits the usefulness of EM, especially in modern applications with complex statistical models (e.g. likelihood inference in random effects models), high-dimensional parameter space (e.g. image reconstruction) and large-scale data (e.g. classification in large databases).

Existing techniques for accelerating the EM algorithm can be divided broadly into two classes: (1) monotone and (2) non-monotone acceleration schemes. The monotone acceleration schemes are intrinsically likelihood-increasing, and hence are stable like the EM algorithm. Examples include efficient data augmentation (Meng & Van Dyk, 1997), parameter expansion EM (PX-EM) (Liu *et al.*, 1998), ECM (Meng & Rubin, 1993) and ECME (Liu & Rubin, 1994). The efficient data augmentation approach, for example, can significantly

increase convergence rates by cleverly augmenting incomplete data using auxiliary working parameters. The non-monotone acceleration methods are essentially root finders, i.e. iterative techniques to find the zeros of some function (e.g. residual of EM fixed-point mapping). They include quasi-Newton (QN) methods (Lange, 1995; Jamshidian & Jennrich, 1997), conjugate gradient (CG) methods (Jamshidian & Jennrich, 1993), and multivariate Aitken's procedure (Louis, 1982).

Despite the availability of numerous acceleration techniques, their use has been limited. We feel that this has to do with the trade-offs involved in the acceleration of the EM algorithm. Important considerations in choosing an accelerator are: (1) generality, (2) ease of implementation, (3) stability, and (4) speed. Monotone acceleration methods such as the efficient data augmentation and PX-EM are not broadly applicable because they require model-specific analytic insights that may not be readily available. Such insights may not even exist for certain classes of problems (e.g. mixture models). Non-monotone acceleration schemes, on the other hand, are broadly applicable to all EM problems. However, their use involves significant trade-offs (Meilijson, 1989). Non-monotone schemes are not intrinsically likelihood increasing, and hence they require good starting values to achieve convergence to a stationary point. They must be 'globalized' in order for them to be practically useful. Globalization involves implementation of safeguards to ensure proper convergence (e.g. checking the negative definiteness of hessian approximation and line search technique for steplength modification). QN methods are prohibitively expensive in high-dimensional problems (e.g. Poisson image reconstruction – Roland *et al.*, 2007) due to storage and handling of approximate hessian matrix. CG methods use minimal storage, but require the evaluation of complete data log-likelihood and its gradient. Multivariate extension of Aitken's procedure, also known as Henrici's or multivariate Steffensen's method (Ortega & Rheinboldt, 1970; Nievergelt, 1991), builds a full-rank approximation of the Jacobian of the EM mapping using forward differences of successive parameter updates. It is seldom used because of numerical instability (Nievergelt, 1991; Jamshidian & Jennrich, 1997).

We develop two new classes of iterative schemes with highly attractive trade-offs for accelerating the EM algorithm. The first, which we call Steffensen-type methods for EM (STEM), is an extension of the scalar Steffensen's method (Henrici, 1964) to vector fixed-point iteration (section 3). The second, which we call 'squared iterative methods' or SQUAREM, is obtained from the STEM using a novel idea called 'squaring' (section 4). SQUAREM is as simple as STEM, but faster and more efficient. Squaring was first suggested by Raydan & Svaiter (2002) to accelerate the convergence of the classical Cauchy (steepest descent) method for solving a linear system of equations. Roland & Varadhan (2005) applied this idea to accelerate convergence of nonlinear fixed-point iterations. Varadhan & Roland (2004) extended these methods by making the connection to vector extrapolation methods with cycling (Smith *et al.*, 1987). Here we further extend Varadhan & Roland (2004) by (a) proposing a new steplength scheme (9), (b) characterizing local convergence behaviour (section 5), and (c) presenting efficient globalization strategies (section 6) for accelerating the convergence of EM.

The proposed methods can be readily implemented as 'off-the-shelf' accelerators, because they (1) require nothing more than the EM parameter updating scheme and (2) are globally convergent. They are especially attractive in high-dimensional problems due to their simplicity and minimal storage requirements, and also in statistical models where model-specific analytic insights are not available to improve EM convergence. We present four examples (Poisson mixture, multivariate t-distribution, causal inference and Monte Carlo EM) to demonstrate their effectiveness (section 7). A pseudocode of SQUAREM listed in Table 1 gives an indication of the simplicity of the proposed schemes. A general-purpose implementation (written in R) is available from the authors.

Table 1. *Pseudocode for SQUAREM*

While not converged
1. $\theta_1 = \text{EMupdate}(\theta_0)$
2. $\theta_2 = \text{EMupdate}(\theta_1)$
3. $r = \theta_1 - \theta_0$
4. $v = (\theta_2 - \theta_1) - r$
5. Compute steplength α (e.g. (9))
6. If necessary, modify α (section 6)
7. $\theta' = \theta_0 - 2\alpha r + \alpha^2 v$
8. $\theta_0 = \text{EMupdate}(\theta')$
9. Check for convergence (23)

2. EM algorithm and its convergence

Let the observations $y = (y_1, \ldots, y_N)$ be generated by a statistical model with the probability density function, $g(y; \theta)$, where $\theta \in \Omega \subset \mathbb{R}^p$. The EM algorithm computes the MLE, θ^*, by iteratively proceeding from a starting value $\theta_0 \in \Omega$. The $(n+1)$th step of the iteration is given as

$$\theta_{n+1} = \text{argmax } Q(\theta; \theta_n); \quad n = 0, 1, \ldots, \tag{1}$$

where

$$Q(\theta; \theta_n) = \int L_c(\theta; x) f(z; y, \theta_n)\, dz, \tag{2}$$

where x is the complete data, $L_c(\theta; x)$ is the complete data log-likelihood and $f(z; y, \theta_n)$ is the density of the missing data z, conditional on y. Computation of the Q-function (2) is the E-step, and the maximization of Q with respect to its first argument is the M-step (1). Wu (1983) provides conditions under which $\{\theta_n\}$ converges to θ^*, which can be either a saddle point or a local maximum.

The EM algorithm implicitly defines a fixed-point mapping F from the parameter space onto itself, i.e. $F : \Omega \subset \mathbb{R}^p \mapsto \Omega$, such that

$$\theta_{n+1} = F(\theta_n), \quad n = 0, 1, \ldots. \tag{3}$$

Under standard regularity conditions, a Taylor series expansion about θ^* yields

$$\theta_{n+1} - \theta^* = J(\theta^*)(\theta_n - \theta^*) + o(\|\theta_n - \theta^*\|), \tag{4}$$

where $J(\theta^*)$ is the Jacobian matrix of $F(\theta)$ evaluated at θ^*. Dempster *et al.* (1977) showed that $J(\theta^*) = I_{\text{miss}}(\theta^*; y) I_{\text{comp}}^{-1}(\theta^*; y)$, where $I_{\text{comp}}(\theta; y) = -\int (\partial^2 L_c(x; \theta)/\partial\theta^2) f(z; y, \theta)\, dz$, and $I_{\text{miss}}(\theta; y) = -\int (\partial^2 \log f(z; y, \theta)/\partial\theta^2) f(z; y, \theta)\, dz$. The Jacobian matrix $J(\theta^*)$ thus measures the fraction of missing information, and the rate of convergence of EM is governed by the largest eigenvalue (in modulus) of $J(\theta^*)$.

3. Steffensen-type methods for EM acceleration

Our goal is to develop faster iterative schemes for locating the fixed point of the EM mapping given by (3), where $F : \Omega \subset \mathbb{R}^p \mapsto \Omega$. Let us consider the sequence of vectors, $\{u_i\}, i \geq 0$, generated by EM starting with $u_0 = \theta_n$. The sequence $\{u_i\}$, is defined as $u_{i+1} = F(u_i), i \geq 0$. Newton's method for finding the fixed point is given by

$$\theta_{n+1} = u_0 - M_n^{-1} g(u_0),$$

where $g(u) = F(u) - u$ is the residual function, $M_n = J(\theta_n) - I$ and $J(\theta_n)$ is the Jacobian of $F(\theta)$ evaluated at θ_n. Newton's method is actually obtained by finding the zero of a linear approximation of $g(\theta)$ around $u_0 : g(\theta) = g(u_0) + M_n \cdot (\theta - u_0)$. For scalar fixed-point problems, Steffensen's method is an attractive alternative to the Newton's method as it provides quadratic convergence without the evaluation of derivatives (Henrici, 1964). As in the scalar case, Steffensen-type methods for vector fixed-point problems can be obtained by building a two-point, secant-like approximation of the Jacobian. We write two different linear approximations for $g(\theta)$ one about θ_n and the other about $F(\theta_n)$:

$$g_0(\theta) = g(u_0) + \frac{1}{\alpha_n} \cdot (\theta - u_0)$$

$$g_1(\theta) = g(u_1) + \frac{1}{\alpha_n} \cdot (\theta - u_1),$$

where we have approximated M_n with the scalar matrix $(1/\alpha_n)I$. The zeros of $g_0(\theta)$ and $g_1(\theta)$ provide two different approximations for the fixed point:

$$t^0_{n+1} = u_0 - \alpha_n g(u_0) \tag{5}$$

$$t^1_{n+1} = u_1 - \alpha_n g(u_1). \tag{6}$$

We now choose α_n such that it minimizes some measure of discrepancy between t^0_{n+1} and t^1_{n+1}. An obvious measure of discrepancy is $\|t^1_{n+1} - t^0_{n+1}\|^2$, leading to the steplength

$$\alpha_n = \frac{r_n^{\mathrm{T}} v_n}{v_n^{\mathrm{T}} v_n}, \tag{7}$$

where $r_n = u_1 - u_0 = F(\theta_n) - \theta_n$, and $v_n = g(u_1) - g(u_0) = F(F(\theta_n)) - 2F(\theta_n) + \theta_n$. The steplength in (7) can be small when $r_n^{\mathrm{T}} v_n$ is small. The Euclidean length $\|t^1_{n+1} - t^0_{n+1}\|$ increases for steplengths greater than that given by (7). Therefore, we also consider other measures that balance discrepancy $\|t^1_{n+1} - t^0_{n+1}\|$ against steplength. One such measure, $\|t^1_{n+1} - t^0_{n+1}\|^2/\alpha_n^2$, yields the steplength

$$\alpha_n = \frac{r_n^{\mathrm{T}} r_n}{r_n^{\mathrm{T}} v_n}. \tag{8}$$

This steplength has the undesirable feature that it can be quite large when $r_n^{\mathrm{T}} v_n$ is small. Minimizing yet another discrepancy measure, $-\|t^1_{n+1} - t^0_{n+1}\|^2/\alpha_n$, by constraining α_n to be negative, yields the steplength:

$$\alpha_n = -\frac{\|r_n\|}{\|v_n\|}. \tag{9}$$

These three choices of steplength, along with (5) provide a new class of iterative schemes for finding the fixed point of EM mapping:

$$\theta_{n+1} = \theta_n - \alpha_n (F(\theta_n) - \theta_n). \tag{10}$$

We call them STEM. We label the scheme in (10) with steplengths given by (7), (8) and (9), as S1, S2 and S3, respectively. The steplength of S3 (9) has desirable properties: (1) it is always negative (independent of the sign of $r_n^{\mathrm{T}} v_n$), (2) it is bounded, and (3) it is the Cauchy–Schwarz lower bound for the steplength in (7) (when $r_n^{\mathrm{T}} v_n < 0$). For a fixed θ_n, we have the relation: $(\alpha_n^{S3})^2 = \alpha_n^{S1} \alpha_n^{S2}$, and when $r_n^{\mathrm{T}} v_n < 0$, we have $\alpha_n^{S2} \le \alpha_n^{S3} \le \alpha_n^{S1} < 0$. These steplengths can also be used along with (6).

We can expand $F(\theta), F^2(\theta)(=F(F(\theta)))$, etc., in Taylor series about the fixed point θ^* and evaluate them at θ_n to obtain:

$$F^j(\theta_n) = \theta^* + J^j \cdot e_n + o(e_n), \quad j = 1, 2, \ldots. \tag{11}$$

where $e_n = \theta_n - \theta^*$ and J is the Jacobian of $F(\theta)$ evaluated at θ^*. From the above identity, we can get the following useful relations:

$$r_n = (J - I) \cdot e_n + o(e_n) \tag{12}$$

$$v_n = (J - I)^2 \cdot e_n + o(e_n). \tag{13}$$

We can also obtain the following recursive error equation that governs the convergence behaviour of STEM in some appropriately small neighbourhood of the fixed-point θ^*:

$$e_{n+1} = [I - \alpha_n(J - I)]e_n + o(e_n). \tag{14}$$

The STEM family of schemes is a simple vector generalization of the classical Steffensen's method for scalar fixed-point problem. Our development makes it clear that they can also be considered as a member of the quasi-Newton family, where the inverse of the Jacobian is approximated by a scalar matrix $\alpha_n I$.

4. Squared iterative methods

Here, we present an exciting new class of schemes which builds on STEM. These new schemes are as simple as STEM, but significantly faster. We motivate the development with a linear problem for two reasons: (1) deeper analytic insights can be gleaned from the linear problem and (2) the fixed point problem posed by EM [cf. (3)] can be well approximated by a linear problem [cf. (4)] in a local neighbourhood of the fixed point.

Let us consider the linear problem:

$$\text{find} \quad x \in \mathbb{R}^p \quad \text{such that} \quad Qx = b, \tag{15}$$

where $Q \in \mathbb{R}^{p \times p}$ is symmetric and positive definite. The classical Cauchy method for this problem is defined by

$$x_{n+1} = x_n - t_n g_n, \quad n = 0, 1, \ldots, \tag{16}$$

where $g_n = Qx_n - b$ and $t_n = g_n^{\mathrm{T}} g_n / g_n^{\mathrm{T}} Q g_n$. Cauchy's method converges linearly, but the rate of convergence can be poor if the smallest eigenvalue of Q is much smaller than the largest one, i.e. if the problem is ill-conditioned. This slow convergence is due to the choice of steplength t_n, but not the gradient direction g_n. To overcome this problem, Barzilai & Borwein (1988) proposed a gradient method (denoted as BB) that can be written as

$$x_{n+1} = x_n - t_{n-1} g_n,$$

where t_{n-1} is the Cauchy steplength at the previous iteration (t_0 is arbitrarily chosen). By combining the Cauchy and BB methods, Raydan & Svaiter (2002) obtained a new method called the Cauchy–Barzilai–Borwein (CBB) method. First, Cauchy scheme (16) is applied at x_n to get an intermediate update, z_n. Then, BB method is applied at z_n, with the previously calculated Cauchy steplength, to obtain the final update, x_{n+1}, as follows:

$$x_{n+1} = x_n - 2t_n g_n + t_n^2 Q g_n.$$

The following recursive error equations hold for the Cauchy and CBB iterative schemes:

$$e_{n+1} = (I - t_n Q)e_n \text{ for Cauchy}, \tag{17}$$

$$e_{n+1} = (I - t_n Q)^2 e_n \text{ for CBB,} \tag{18}$$

where $e_n = x_n - x$. Based on (17) and (18), we consider CBB method as a 'squared' Cauchy method.

Raydan & Svaiter (2002) demonstrated the superiority of CBB scheme over the Cauchy and BB methods. Therefore, it seems natural to adapt this idea to solve the nonlinear fixed-point problem of EM. We obtain the new class of schemes by demanding that they satisfy the following recursive error relation obtained by squaring (14) analogous to the relation between (17) and (18):

$$e_{n+1} = [I - \alpha_n (J - I)]^2 e_n. \tag{19}$$

Using the identities (12) and (13) we obtain the new scheme as

$$\theta_{n+1} = \theta_n - 2\alpha_n r_n + \alpha_n^2 v_n. \tag{20}$$

We will call these new methods as 'SQUAREM'. The SQUAREM schemes with steplengths given by (7), (8) and (9) will be denoted as SqS1, SqS2 and SqS3, respectively.

5. Local convergence

Roland & Varadhan (2005) gave a proof of convergence for SQUAREM under restrictive conditions on the EM mapping $F(\theta)$, but that proof does not provide any insight on convergence behaviour. To gain such insight, we consider the linear problem (15), which is equivalent to finding the fixed point $F(x) = x$, with $F(x) = (I + Q)x + b$. Thus, in this particular case, $r_n = Qx_n - b$ and $v_n = Qr_n$ in (7), (8) and (9). As any positive definite matrix can be diagonalized via an orthogonal transformation, without any loss of generality we will let Q be a diagonal matrix with positive entries $1/\lambda_i, i = 1, \ldots, p$ such that $0 < \lambda_1 \leq \cdots \leq \lambda_p$. The recursive error relation for the ith component of error (for STEM) can be written as

$$e_{n+1}^{(i)} = \left[I - \frac{\alpha_n}{\lambda_i} \right] e_n^{(i)}, \quad i = 1, \ldots, p. \tag{21}$$

We can choose α_n from (7), (8) or (9), with $r_n = Qx_n - b$ and $v_n = Qr_n$. In fact, S2 corresponds to the classical Cauchy scheme. As $r_n = Q^{-1} v_n$, the α_n from the three schemes can be written as a Rayleigh quotient. For example, for S1 scheme we have

$$\alpha_n = \frac{v_n^{\mathrm{T}} Q^{-1} v_n}{v_n^{\mathrm{T}} v_n}.$$

Using the extremal properties of Rayleigh quotient principle (Lancaster, 1969), we have $\lambda_1 < \alpha_n < \lambda_p$. In conjunction with (21), this means that the error components corresponding to the smallest and largest eigenvalues (hereafter we will simply call these as the smallest and largest error components) will never be totally annihilated, but when $\alpha_n = \lambda_i, i \neq 1, p$, the ith error component will be annihilated at the nth iteration, and will remain so thereafter. Typically, the steplength α_n will be closer to λ_1 in the initial iterations. This attenuates the overall error, and in particular, the small error components are significantly diminished. But the large error components will not be attenuated appreciably. Therefore, the steplength is increased in subsequent iterations. On the other hand, when α_n is closer to λ_p, the small error components will be magnified, causing the scheme to take smaller steps on subsequent iterations. This results in a characteristic oscillatory pattern of steplengths. The new schemes will exhibit a monotonic decrease in error when the steplength oscillates between smaller eigenvalues in the initial iterations, with the amplitude of oscillation becoming larger as convergence is approached. The overall error can even increase if the scheme takes a large step during the initial iterations before the small error components have decreased significantly. This,

generally, is not deleterious, since monotonicity is not necessary for convergence. It is, however, critically important that the steplengths be selected from all parts of the interior of the spectrum of eigenvalues in order that each individual error component can be decreased significantly as dictated by (21).

We illustrate this convergence behaviour with a simple three-dimensional linear system, where $\boldsymbol{\lambda} = (0.1, 1, 100)$ and $b = (-10, -100, 0.1)$. The solution is $x = (-1, -100, 10)$. We use an initial value $x_0 = (0, 0, 0)$. STEM schemes are extremely slow, with S1 and S2 taking nearly 3200 and 1800 iterations to get within 10^{-7} of the solution. S3 only came within 4×10^{-4} of the solution after 5000 iterations. The slow convergence of STEM was due to their inability to attenuate the error corresponding to $\lambda_3 = 100$. Their largest steplength was 3 orders of magnitude smaller than λ_3. In contrast, all three SQUAREM schemes converged very fast (SqS1 = 8, SqS2 = 10, SqS3 = 9 iterations) to within 10^{-7} of the solution. Figure 1 shows the error $\|x_n - x\|$ and steplength α_n for the SqS1 scheme (SqS2 and SqS3 behave similarly). We can see the steplength (depicted in dashed lines) oscillating between the smallest and largest eigenvalues 0.1 and 100 (depicted in dotted lines) with increasing amplitude. During the course of oscillation, steplengths of the SQUAREM get very close to all three eigenvalues, thus diminishing all the error components. Overall, we can also observe the monotone, superlinear decrease in error, although, more precisely, convergence occurs at different linear rates between cycles, with increasingly faster rates closer to solution.

This type of convergence is quite typical even in nonlinear EM iterations, where the 'local' rates of convergence of STEM and SQUAREM are governed by the spectral radius of $I - \alpha_n(J - I)$. We assume that the eigenvalues of $J - I$ lie in $[-1, 0)$ (this holds when the observed information matrix is positive definite). Using the extremal properties of Rayleigh quotient, it can be shown that the steplengths of S1, S2 and S3 satisfy the following relation in some neighbourhood of the solution: $-\infty < \lambda_p < \alpha_n < \lambda_1 < -1$, where λ_1 and λ_p are the smallest and largest (negative) eigenvalues of $(J - I)^{-1}$. Interestingly, the greater the ill-conditioning of the problem (λ_p/λ_1), the greater the acceleration of EM by SQUAREM and STEM.

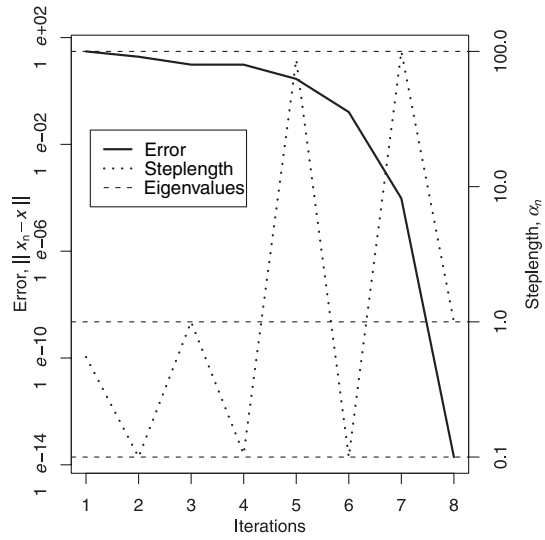

Fig. 1. Convergence behaviour of SqS1 scheme for a simple, three-dimensional linear problem. Errors (in log scale) are on the left Y-axis and steplengths (in log scale) are on the right Y-axis. Three eigenvalues are shown as dotted lines and their values are on the right Y-axis.

6. Global convergence

The analysis in section 5 is strictly valid only in a small neighbourhood of the fixed point. It provides no assurance about how the steplengths behave when the starting parameter value is 'far' from the fixed point. In fact, the steplengths given by (7) and (8) can even be positive for poor starting values, which is generally undesirable. Therefore, we need a 'globalization' strategy that modifies the steplengths, if necessary. A globally convergent method converges to a stationary point from any starting point in the parameter space, or at least, in a large part of it. The stationary point can either be a saddle point or a local maximum that is not necessarily a global maximum (see Dennis & Schnabel, 1983, pp. 4–5, for a good description of global convergence). We can obtain globally convergent STEM and SQUAREM by modifying the steplengths such that the parameter update increases the observed-data log-likelihood. The important question is whether we can always find a steplength $\alpha_n < 0$, at any parameter θ_n. We show that we can even do better by finding a steplength $\alpha_n < -1$. Here is a simple proof.

At any point θ_n in the interior of the parameter space, let us define the SQUAREM update as a function of steplength α:

$$\theta_{n+1}(\alpha) = \theta_n - 2\alpha r_n + \alpha^2 v_n, \tag{22}$$

where $r_n = F(\theta_n) - \theta_n$ and $v_n = F(F(\theta_n)) - 2F(\theta_n) + \theta_n$. This is a continuous function of α, which can be viewed as a continuation of the EM algorithm, as $\theta_{n+1}(-1) = F(F(\theta_n))$ (similarly, the STEM update is also a continuation of EM, as $\theta_{n+1}(-1) = F(\theta_n)$). Now, the likelihood update is given by: $L(\alpha) := L(\theta_{n+1}(\alpha))$. For any given θ_n, this is a continuous function of α (if we assume that the likelihood is continuous in parameters). By the ascent property of EM, we have $L(-1) > L(0)$. As $L(\alpha)$ is continuous, there exists an open interval around -1, i.e. $I_\alpha = (\alpha^*, -1)$, such that for any $\alpha \in I_\alpha$, $L(\alpha) > L(0)$. This completes the proof.

To devise an efficient computational strategy that ensures monotonicity, we can use the continuation parameter α as the currency to trade-off speed (efficient reduction of all error components) for stability (monotonicity of observed-data likelihood). If we stick 'close' to EM ($\alpha \approx -1$) we should inherit the stability of EM, but this will not improve the rate of convergence significantly. Whenever feasible, we should take the steplength, α_n, given in (9). This ensures that we achieve a greater reduction of the error component corresponding to the largest (negative) eigenvalue. Here is our strategy: If $\alpha_n > -1$, we set $\alpha_n = -1$. If $\alpha_n < -1$, we consider two cases: (1) if $L(\alpha_n) \geq L(0)$, we will accept α_n as our steplength, (2) else if $L(\alpha_n) < L(0)$, we will successively increase α_n towards -1 until it belongs to I_α. This strategy is called back-tracking. A simple back-tracking strategy is to repeatedly halve the distance between α_n and -1: $\alpha_n \leftarrow (\alpha_n - 1)/2$, until $L(\alpha_n) > L(0)$.

Global convergence can be rigorously proved using the device of Lyapunov functions for discrete dynamical systems (Ortega, 1973), based primarily on the following facts: (1) the observed-data log-likelihood is a Lyapunov function and (2) SQUAREM iteration (with steplength modification) is continuous in parameter θ. The globally convergent STEM (gSTEM) and SQUAREM (gSQUAREM) schemes will be denoted as gS1, gS2, gS3 and gSqS1, gSqS2 and gSqS3, respectively.

7. Test problems and results

We demonstrate the performance of the new methods on four statistical problems. The stopping criterion is based on the residual error,

$$\|F(\theta_n) - \theta_n\| \leq \varepsilon, \tag{23}$$

where we chose $\varepsilon = 10^{-7}$. A maximum of 10,000 fevals is allowed to achieve convergence. We used three performance measures to evaluate and compare different schemes: total CPU time in seconds, number of EM evaluations (*Fevals*) and number of times a scheme failed. A scheme can fail in one of three ways: exceeding 10,000 EM evaluations, violating parameter constraints or converging to a non-optimal solution (smaller likelihood than that of the best scheme). We report both the mean and the 95% probability interval (the 2.5th and 97.5th percentiles for CPU times), *Fevals* and *Levals* (number of log-likelihood evaluations by globally convergent schemes). The number of EM evaluations measures the rate of convergence of the schemes, whereas the total CPU time is a measure of the overall computational efficiency.

There are 12 numerical schemes to be evaluated in relation to the EM as per the design: (STEM/SQUAREM) $\times$ (S1/S2/S3) $\times$ (globally convergent/non-monotonic). We have performed extensive evaluations of all the 12 schemes on a number of test problems, including the four problems described below. These numerical experiments clearly demonstrated the superiority of SQUAREM compared with STEM. The two SQUAREM schemes SqS1 and SqS3, with steplengths α_n^{S1} (7) and α_n^{S3} (9), were the picks of the lot, although SqS3 had a slight edge in terms of being consistently faster than SqS1. They both were better than SqS2. For the sake of simplicity and to avoid clutter, we only present the results of SqS3 and gSqS3. In particular, we evaluated (1) the stability and computational efficiency of SqS3 and gSqS3 (globally convergent SqS3) compared with the EM algorithm, (2) the computational efficiency of gSqS3 relative to SqS3, and (3) the trade-off between stability and computational efficiency via comparison between gSqS3 and SqS3. We did not perform any likelihood computations for the EM and SqS3 schemes in order to make a fair, but conservative, comparison with gSqS3, as likelihood evaluation is not necessary for EM and SqS3. All the computations were performed in R, Version 2.3.1 (R Development Core Team, 2006), on a Windows platform using a 3.4 GHz Pentium 4 processor with 1GB RAM.

7.1. Poisson mixtures

We consider the famous data from *The London Times* during the years 1910–1912 reporting the number of deaths of women 80 years and older (Hasselblad, 1969). The tabulation was in terms of the number of days, n_i, in which i deaths occurred (see Table 2). A two-component mixture of Poisson distribution provides a good fit to the data, whereas a single Poisson distribution had a poor fit.

We chose this simple example primarily to illustrate the local-convergence properties of SQUAREM. The likelihood for this problem can be written as

$$\prod_{i=0}^{9} \left[p e^{-\mu_1} \mu_1^i / i! + (1-p)\, \mathrm{e}^{-\mu_2} \mu_2^i / i! \right]^{n_i}.$$

We have to estimate three parameters, $\theta = (p, \mu_1, \mu_2)$. The EM algorithm is as follows:

Table 2. *Data from* The London Times *on deaths during 1910–1912*

Deaths, y_i	Frequency, n_i	Deaths, y_i	Frequency, n_i
0	162	5	61
1	267	6	27
2	271	7	8
3	185	8	3
4	111	9	1

$$p^{(k+1)} = \sum_i n_i \hat{\pi}_{i1}^{(k)} \Big/ \sum_i n_i,$$

$$\mu_j^{(k+1)} = \sum_i i n_i \hat{\pi}_{ij}^{(k)} \Big/ \sum_i n_i \hat{\pi}_{ij}^{(k)}, \quad j = 1, 2,$$

where $\hat{\pi}_{ij}^{(k)} = p^{(k)} \left(\mu_j^{(k)}\right)^i \mathrm{e}^{-\mu_j^{(k)}} \Big/ \left[p^{(k)} \left(\mu_1^{(k)}\right)^i \mathrm{e}^{-\mu_1^{(k)}} + (1 - p^{(k)}) \left(\mu_2^{(k)}\right)^i \mathrm{e}^{-\mu_2^{(k)}}\right], \quad j = 1, 2.$

The MLEs for the parameters are: $(\hat{p}, \hat{\mu_1}, \hat{\mu_2}) = (0.3599, 1.256, 2.663)$. The EM is very slow to converge for this problem, regardless of the starting value, because the data does not contain adequate information to clearly separate the two components. The eigenvalues of the Jacobian of the EM mapping at the MLE, $J(\theta^*)$, are 0, 0.7204 and 0.9957, and those of $(J - I)^{-1}$ are -1.00, -3.58 and -230.7. The slow convergence of the EM is due to the largest (negative) eigenvalue.

We report the performance of the schemes for 5000 randomly chosen starting values: $\theta_0 = (p_0, \mu_{10}, \mu_{20}) = [U(0.05, 0.95), U(0, 100), U(0, 100)]$, where $U(a, b)$ is a uniform random deviate between a and b. These starting values should provide a stringent test of the global convergence of SQUAREM, in relation to EM. Table 3 reports the results. Both SqS3 and gSqS3 provided major acceleration of EM. SqS3 was faster than EM by a factor of more than 16 in terms of CPU time and by a factor of 25 in terms of rate of convergence. The monotonic gSqS3 had the same acceleration of the rate of convergence of EM as SqS3, but was a bit slower than SqS3, in terms of CPU time, due to observed-data likelihood evaluations. However, gSqS3 was superior in terms of stability as it never failed, just like the EM. SqS3 had 169 failures: 121 due to convergence to a degenerate, single component solution, and 48 due to violation of parameter constraints.

Figure 2 depicts the convergence behaviour of SqS3 for the starting value: $\theta_0 = (0.3, 1.0, 2.5)$. We notice that the steplength oscillates between the eigenvalue spectrum, -3.58 and -230. The amplitude of oscillation increases gradually as convergence is approached. Every cycle of oscillation in steplength (going from -3.6 to -230, and then back to -3.6), with increasing amplitude, results in a major reduction in overall error. Overall convergence is monotone and superlinear.

7.2. Multivariate t-distribution

Here, we demonstrate that SQUAREM can improve the convergence of EM and PX-EM in multivariate t-distribution. Our main goal in this example is to illustrate the generality of SQUAREM for accelerating any generalized EM (GEM) algorithm. This is especially useful in problems where problem-specific analytic insights, such as those required for PX-EM, are not available, but only the slowly converging GEM schemes (e.g. ECM, ECME and gradient EM) can be implemented. Meng & Van Dyk (1997) used the multivariate t-distribution example to demonstrate efficient data augmentation. Although they did not give a specific name to

Table 3. *Poisson mixture estimation with 5000 random starting values*

Scheme	No. fevals	No. levals	No. failures	CPU (s)
EM	2386 (2267, 2571)	0	0	0.35 (0.33, 0.37)
SqS3	94 (72, 102)	0	169	0.022 (0.01, 0.03)
gSqS3	94 (81, 108)	68 (59, 78)	0	0.036 (0.03, 0.05)

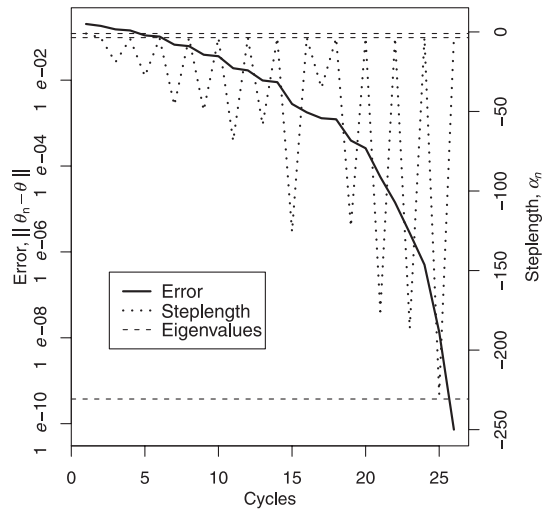

Fig. 2. Convergence behaviour of SqS3 scheme for Poisson mixture problem. Errors are on the left Y-axis and steplengths (in dotted lines) are on the right Y-axis. Three eigenvalues (-1.00, -3.58, -230.8) are shown as dashed lines and their values are on the right Y-axis.

their method, we shall call it PX-EM, as it fits within the 'parameter expansion' framework proposed later in Liu *et al.* (1998) (see the discussion section of this paper for a subtle distinction between PX-EM and Meng & Van Dyk's efficient data augmentation).

The multivariate t-distribution is commonly used for robust estimation. Its density is

$$f_v(y \mid \mu, \Sigma) \propto |\Sigma|^{-1/2}\{v + (y-\mu)^{\mathrm{T}}\Sigma^{-1}(y-\mu)\}^{-(v+p)/2}, \quad y \in \mathbb{R}^p.$$

Fitting this model to observed data $y = (y_1, \ldots, y_N)$ requires maximizing the likelihood function $\prod_i f_v(y_i \mid \mu, \Sigma)$, over (μ, Σ), for a known degree of freedom v. There is no general closed form solution for the MLE, but if we augment y such that $Y_{\mathrm{comp}} = \{(y_i, q_i), i = 1, \ldots, N\}$, where q_i are i.i.d. from χ^2_v/v, then the MLE follows from a weighted least-squares procedure. The E-step finds the expectation of the complete-data log-likelihood, conditionally on y and (μ_n, Σ_n) from the previous iteration. The missing data q_i, conditionally on y, and (μ_n, Σ_n) is distributed as: $q_i \mid y \sim \chi^2_{v+p}/(v + d_i^{(n)})$, where $d_i^{(n)} = (y_i - \mu_n)^{\mathrm{T}}\Sigma_n^{-1}(y_i - \mu_n)$, $i = 1, \ldots, N$. As the complete-data log-likelihood is linear in q_i, the E-step is:

$$w_i = E(q_i \mid y_i, \mu_n, \Sigma_n) = (v+p)/(v + d_i^{(n)}), \quad i = 1, \ldots, N.$$

Using these weights in the weighted least-squares procedure yields the M-step:

$$\mu_{n+1} = \sum_i w_i y_i \Big/ \sum_i w_i,$$

$$\Sigma_{n+1} = \frac{1}{N}\sum_i w_i (y_i - \mu_n)(y_i - \mu_n)^{\mathrm{T}}. \tag{24}$$

The PX-EM of Meng & Van Dyk (with optimal working parameter) is trivially different from the standard EM – replace the denominator N in (24) with the sum of the weights w_i:

$$\Sigma_{n+1} = \sum_i w_i (y_i - \mu_n)(y_i - \mu_n)^{\mathrm{T}} \Big/ \sum_i w_i.$$

Meng & Van Dyk showed that the PX-EM, although still linearly convergent, converges faster than the standard EM, while maintaining the likelihood ascent property. It can be

Table 4. *Parameter estimation of a 10-dimensional multivariate t-distribution, with 1 degree of freedom with $\mu = 0$, and a randomly generated covariance matrix. Results of 5000 simulations with data-based initial values from (25)*

Method	No. Fevals	No. Levals	No. Failures	CPU (s)
EM	233 (218, 250)	0	0	19.0 (17.7, 21.1)
SqS3	47 (39, 57)	0	0	3.9 (3.2, 4.9)
gSqS3	50 (39, 63)	36 (27,46)	0	6.3 (4.8, 8.1)
PXEM	22 (20, 25)	0	0	1.9 (1.7, 2.2)
SqS3-PXEM	18 (18, 21)	0	0	1.5 (1.4, 1.7)
gSqS3-PXEM	18 (18, 21)	13 (13,15)	0	2.2 (2.2, 2.6)

verified that the largest eigenvalue of the Jacobian of the optimal PX-EM mapping is the same as the second largest eigenvalue of the Jacobian of EM mapping (a fact not mentioned by Meng & Van Dyk), explaining its faster convergence.

We report the results of 5000 simulations of fitting a 10-dimensional ($p = 10$) Cauchy model ($v = 1$) to 100 observations generated from $f_1(y \mid \mu = 0, \Sigma = V)$, where V is an arbitrary variance–covariance matrix (symmetric and positive definite) that was randomly determined at the outset of the simulation. These simulations are the same as conducted by Meng & Van Dyk. There are 65 parameters (10 for μ and 55 for Σ) to be estimated. We used initial values suggested by Meng & Van Dyk:

$$\mu_0 = \bar{y}; \quad \Sigma_0 = \left(\frac{1}{N}\right) \sum_i (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}}. \tag{25}$$

Table 4 presents the results. SqS3 was faster than EM by a factor of 5. Even though gSqS3 had the same acceleration of rate of convergence as SqS3, it was less efficient with only a three-fold reduction in CPU time. PX-EM provides an amazing 10-fold acceleration of EM (although the acceleration is less dramatic for $v > 1$). Even though SQUAREM has a fast, superlinear rate of convergence, taking only around 15 cycles to converge, the overall computational effort is larger than that of PX-EM, as SQUAREM computes three EM updates in each cycle. It is important to reiterate that even though PX-EM provides substantial gains over EM, it requires model-specific analytic insights. SQUAREM, on the other hand, is broadly applicable in any EM problem.

PX-EM, although very fast, is still only linearly convergent. So, we ask whether we can further improve its rate of convergence using SQUAREM. Applying the SQUAREM scheme, SqS3, to PX-EM accelerates it by 20%, and provides an overall 13-fold acceleration over standard EM. The monotonic gSqS3 also improves the rate of convergence of PX-EM by 20% in terms of the number of Fevals, but it takes more CPU time due to likelihood evaluations.

7.3. Principal stratification models in causal inference

The increasing attention on causal relations in diverse fields such as public health and economics, has spurred the study of effects of treatments that are only partially controlled. Estimation of causal effects in such partially controlled studies can be undertaken using the general framework of 'principal stratification' (Frangakis & Rubin, 2002). In relation to our discussion, this framework gives rise to a likelihood which is a mixture over a partially unobserved stratification. To demonstrate the performance of our acceleration methods in such likelihood structure, we use an example from the evaluation of the impact of Baltimore's needle exchange program (NEP) in reducing the HIV transmission among injection drug users (Frangakis *et al.*, 2004). In this study, the investigators determined the location

of the NEP sites, to which the injection drug users could go and exchange on a one-to-one basis used needles for new, sterile ones. Generally, proximity to the NEP site affects both exchange behaviour and the provision of serum samples for monitoring HIV status. Therefore, the controlled location of the NEP sites can affect the measures of interest, the exchange behaviour and the ability to ascertain HIV status, which are not directly controlled.

To define the effect of exchange on HIV, Frangakis *et al.* (2004) considered the stratification of each subject i by their exchange behaviour, $S_i = (E_i(1), E_i(0))$, where $E_i(1)$ is a binary indicator of exchange behaviour if the NEP site is placed near the subject's residence, and $E_i(0)$ denotes exchange behaviour if the NEP site is placed far. Then, the effect that exchanging needle has on HIV can be defined as the effect that controlled distance (near or far) has on HIV among the principal stratum $S_i = (0, 1)$. To estimate this effect, the following three-component likelihood model was used: (1) a model, $f^{(y)}(Y_i \mid D_i, H_i, S_i; \beta^{(y)})$, for the outcome variable Y_i, the HIV status of the unit contingent upon the unit being at risk, in terms of distance D_i, past history H_i and principal stratum S_i, (2) a model, $f^{(c)}(C_i \mid D_i, H_i, S_i; \beta^{(c)})$, for the censoring indicator C_i, i.e. whether or not a unit provides serum samples for outcome ascertainment, and (3) a model, $f^{(s)}(S_i \mid H_i; \beta^{(s)})$, for the principal stratum S_i. The observed data for a unit i are $(H_i, D_i, E_i, C_i, C_i Y_i)$. The likelihood for the observed data can be written as:

$$L_{\text{obs}}(\boldsymbol{\beta}) = \prod_i \sum_{s_i} \Pr(S_i = s \mid H_i, D_i, \beta^{(s)}) \times \Pr(C_i \mid S_i = s, H_i, D_i, \beta^{(c)}) \times \Pr(Y_i \mid S_i = s_i, H_i, D_i, \beta^{(y)})^{C_i},$$

where s_i is the set of principal strata for which the exchange status is E_i when the distance is D_i and $\boldsymbol{\beta} = (\beta^{(s)}, \beta^{(c)}, \beta^{(y)})$ is the vector of all the parameters. We see that the principal stratification model is a mixture model, where the weights are determined by the distribution of the latent principal strata. We can write the likelihood for the complete data (which is observed data + S_i) as

$$L_c(\boldsymbol{\beta}) = \prod_{s_i} g_i(s; \boldsymbol{\beta})^{\mathbf{1}\{s_i = s\}},$$

where

$$g_i(s; \boldsymbol{\beta}) = f^{(s)}(s \mid H_i, \beta^{(s)})\, f^{(c)}(C_i \mid D_i, S_i = s, H_i, \beta^{(c)}) \left[f^{(y)}(Y_i \mid H_i, D_i, S_i = s, \beta^{(y)})\right]^{C_i},$$

from which we obtain the distribution of the principal stratum conditional on observed data as

$$\pi_i(s, \boldsymbol{\beta}) = \Pr(S_i = s \mid \text{observed data}) = \frac{g_i(s; \boldsymbol{\beta})}{\sum_{s_i} g_i(s; \boldsymbol{\beta})} \mathbf{1}\{s \in s_i\}.$$

Now we can compute the E-step, at the $(m+1)$th iteration, as the expectation of L_c conditioned on the observed data, as

$$E\left[\log L_c \mid \text{observed data}\right] = \sum_s \pi_i(s, \boldsymbol{\beta}_m) \Big[\log f^{(s)}(s \mid H_i, \beta^{(s)}) + \log f^{(c)}(C_i \mid D_i, s, H_i, \beta^{(c)}) + C_i \log f^{(y)}(Y_i \mid D_i, s, H_i, \beta^{(y)})\Big],$$

where $\boldsymbol{\beta}_m$ is the vector of parameter values from the mth iteration. This function needs to be maximized with respect to $\boldsymbol{\beta} = (\beta^{(s)}, \beta^{(c)}, \beta^{(y)})$ to obtain the parameters for the next iteration. This is the M-step, which is easily performed by noting that the maximization simplifies to three separate maximizations corresponding to the models for S, C and Y, using as weights, $\pi_i(s, \boldsymbol{\beta})$, for each principal stratum.

Table 5. *Acceleration of EM algorithm for the principal stratification models for the NEP evaluation problem*

	EM	SqS3	gSqS3
CPU(s)	1718	290	352
Fevals	272	45	54
Log-likelihood	−5461.851	−5461.851	−5461.851

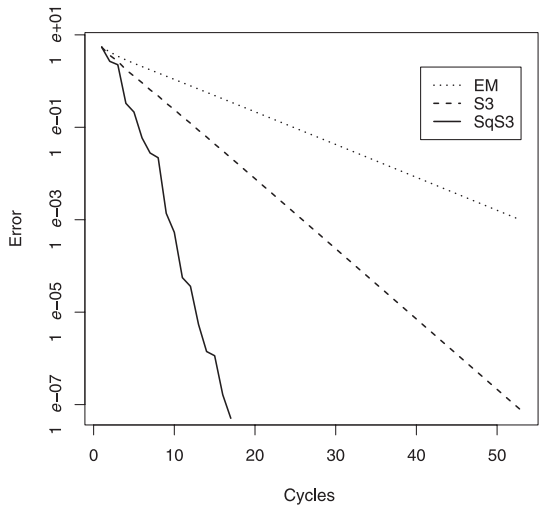

Fig. 3. Convergence behaviour of schemes for principal stratification model. Note that each cycle of S3 and SqS3 involves three EM updates, therefore, the errors corresponding to every third iteration (i.e. $3, 6, \ldots$) of EM are plotted.

The details of the study design and data characteristics are available in Frangakis *et al.* (2004). R functions implementing the EM algorithm for this principal stratification model are provided in a software package called 'PSpack' (Frangakis & Varadhan, 2004). The results of the EM algorithm are presented in Table 5. Both the E and the M steps are time consuming for this problem. EM took nearly 30 min to converge as per (23). SQUAREM (SqS3) converged in around 5 min, accelerating the EM by a factor of 5–6. The globally convergent gSqS3 scheme was slightly slower than SqS3, as it took two more cycles to converge. Figure 3 provides a picture of the rate of convergence of EM, S3 (result not shown in Table 5) and SqS3. S3 has a faster, linear convergence than the EM, whereas SqS3 exhibits superlinear convergence.

7.4. Improving the convergence of Monte Carlo EM

Here, we explore whether SQUAREM can improve the convergence of EM in cases where the E-step is analytically intractable and hence is evaluated using Monte Carlo-type simulations. This is known as the Monte Carlo EM (MCEM) algorithm. Here we consider a generalized linear mixed model (GLMM) example, a binomial model with logit link and a normally distributed random effect. This is a standard benchmark problem in the MCEM literature (McCulloch, 1997; Booth & Hobert, 1999). The model is as follows:

$$
\begin{aligned}
&Y_{ij} \sim \text{indep Bernoulli}(p_{ij}), \quad i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, M \\
&\log(p_{ij}/(1-p_{ij})) = \beta x_{ij} + u_j \\
&u_j \sim \text{i.i.d.} N(0, \sigma^2).
\end{aligned} \tag{26}
$$

We took $N = 15$, $M = 10$, $x_{ij} = i/15$, and simulated two data sets from model (26), one for small and the other for large random effect variance, σ^2. The MLE of β and σ^2 for these two cases, computed by directly maximizing the likelihood for response Y (after integrating out the random effect using a 50-point Gauss–Hermite quadrature), were:

$$
\begin{aligned}
&\hat{\beta} = 1.699, \hat{\sigma}^2 = 0.0534 \quad \text{(small variance)}, \\
&\hat{\beta} = 2.081, \hat{\sigma}^2 = 1.410 \quad \text{(large variance)}.
\end{aligned}
$$

To compute the E-step of the MCEM algorithm, we use a rejection sampling scheme that produces independent draws of the random effect u (Booth & Hobert, 1999).

The main challenge in the use of MCEM and SQUAREM here is due to the error and the Monte Carlo (MC) variance associated with the E-step. Consequently, the EM mapping $F(\cdot)$ is not smooth. In the initial stages of MCEM, when the starting values are generally far from the solution, EM takes larger steps and a relatively small MC sample size would suffice. As EM approaches the solution, the sample size needs to be increased appropriately so that the sampling error does not overwhelm the EM step. How to adequately and efficiently increase the sample size to ensure proper convergence is an important research problem in the MCEM literature (Booth & Hobert, 1999; Caffo *et al.*, 2005). Here we eschewed complications involving adaptive sampling and determined the sample size for each successive E-step at a geometrically increasing rate as follows: $\{n_1, n_1, n_1, n_2, n_2, n_2, \ldots, n_k, n_k, n_k, \ldots, n_K, n_K, n_K\}$, where $n_1 = 100, n_K = 5000,$ and $n_{k+1} = n_k * \delta, k = 1, 2, \ldots, K-1,$ with $\delta = (n_K/n_1)^{1/(K-1)}$. We kept the same MC sample size for three consecutive E-steps in order to have same computational effort for both MCEM and SQUAREM (which computes three MCEM steps in each cycle).

We took $K = 10$, and ran 100 simulations each for the small and large variance scenarios, with the same starting value ($\beta_0 = 1, \sigma_0^2 = 1$). We then compared the accuracy of the solutions at the end of 30 MCEM steps (or 10 cycles of SQUAREM) by using the number of significant digits in the solution compared to the true MLEs, defined as: $-\log_{10}((\|\theta_K - \theta^*\|)/\|\theta^*\|)$, where θ_K is the solution from MCEM or SQUAREM, and θ^* is the MLE.

Results are shown in Table 6 and Figs 4 and 5. In the small variance case, SQUAREM provided significantly more accurate solutions by accelerating the convergence of MCEM. MCEM was more accurate than SQUAREM only in four of 100 simulations (Fig. 4). The top panel of Fig. 5 clearly shows the faster rate of convergence of SQUAREM for this case. In the large variance case, however, SQUAREM does not provide any acceleration of MCEM. This is mainly because, the MCEM algorithm itself converges rapidly and hits the

Table 6. *Accuracy of the solutions provided by MCEM and SQUAREM in terms of number of significant digits that are present in the true MLE*

	MCEM	SQUAREM
Small variance	1.70 (1.65–1.73)	2.10 (1.82–2.30)
Large variance	2.22 (1.99–2.38)	2.17 (1.95–2.34)

Values reported are means over 100 simulations, with inter-quartile ranges in parentheses.

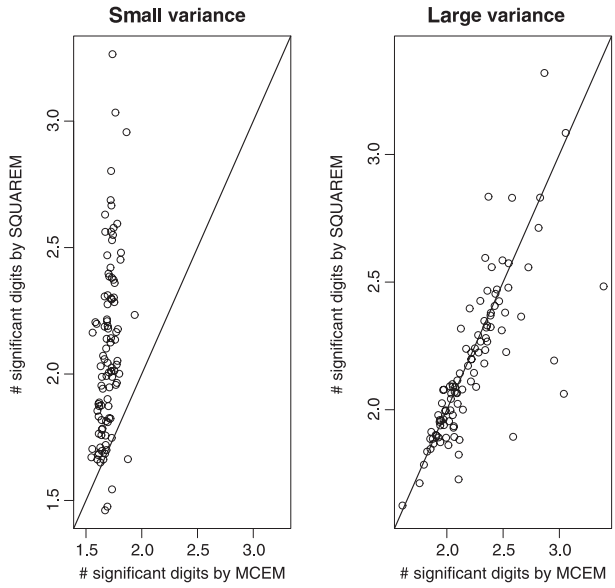

Fig. 4. Comparison of the accuracy of solutions by MCEM and SQUAREM for small (left) and large (right) variance case.

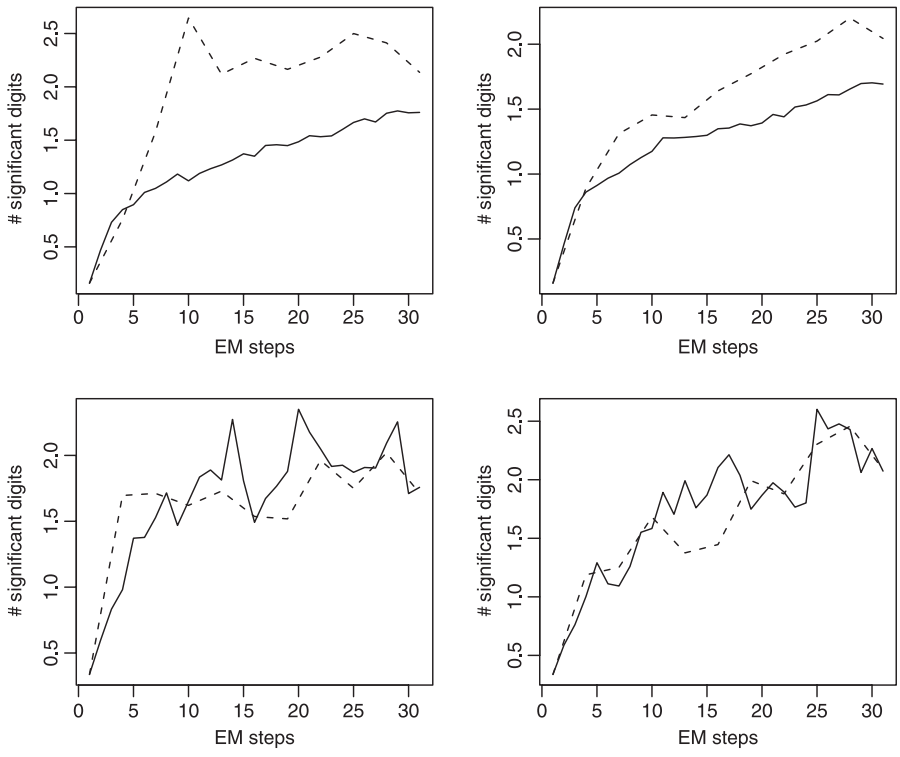

Fig. 5. Comparison of the rates of convergence of MCEM and SQUAREM for small (top two panels) and large (bottom two panels) variance cases. Solid and dashed lines, respectively, denote MCEM and SQUAREM.

Monte Carlo noise barrier quite fast (around 10 MCEM steps), as can be seen from the bottom panel of Fig. 5.

8. Discussion

We have presented a new class of iterative methods, called SQUAREM, for accelerating the EM algorithm. These methods are based on the novel idea of 'squaring' applied to another new class of methods called STEM. Like quasi-Newton and conjugate gradients, SQUAREM is non-monotone in observed-data likelihood. However, there is an important difference. Elaborate procedures are required to ensure global convergence for QN and CG (even then these schemes can fail). On the other hand, as shown in section 6, the continuation property of SQUAREM enables a simple steplength backtracking strategy to obtain global convergence. It was proved in section 6 that this strategy will always succeed. The only added complexity in the globally convergent SQUAREM is the need to evaluate observed-data likelihood, at least twice, in each cycle.

Our empirical results suggest that SQUAREM can achieve superlinear convergence rates, especially when there is a large fraction of missing data. In particular, SQUAREM achieves rapid, superlinear convergence when there are a few large (negative) eigenvalues of $(J - I)^{-1}$ that are well separated from most other eigenvalues close to -1. Thus, SQUAREM is most effective when there is a greatest need to accelerate the EM.

SQUAREM is especially attractive in high-dimensional problems (e.g. PET image reconstruction – Roland *et al.*, 2007) when compared with numerical accelerators such as quasi-Newton and conjugate gradient methods, because of its simplicity and minimal storage requirements. SQUAREM can improve the rate of convergence of any generalized EM (GEM) algorithm (e.g. ECM, ECME and gradient EM). This is especially useful in problems where model-specific analytic insights, such as those required for efficient data augmentation and PX-EM, are not available, but only the slowly converging GEM schemes can be implemented. SQUAREM can also accelerate the MM (minorize and maximize) algorithms, which do not invoke the missing-data framework, but are monotone like the EM (Lange *et al.*, 2000). In fact, SQUAREM also shows promise for improving the convergence and efficiency of Monte Carlo EM. Further evaluations of SQUAREM are needed, especially in conjunction with more sophisticated adaptive sampling approaches such as the ascent-based Monte Carlo EM (Caffo *et al.*, 2005).

In summary, SQUAREM has an important advantage that none of the existing methods of acceleration possess, namely, that it can be readily implemented as an 'off-the-shelf' accelerator of any EM-type algorithm, because it (1) only requires EM parameter updating scheme, (2) does not require auxiliary quantities such as gradient and hessian of log-likelihood, (3) uses only $O(4p)$ storage, where p is the number of parameters, and (4) is globally convergent. The pseudocode given in Table 1 provides an indication of the simplicity of SQUAREM. A general-purpose implementation of SQUAREM (written in R) is available from the authors.

Acknowledgements

The authors thank Dr Claude Brezinski for initiating their intercontinental collaboration. Special thanks go to Drs Constantine Frangakis and Tom Louis for the many useful discussions. Drs Brian Caffo, Hormuzd Katki and Marcos Raydan gave valuable feedback on the manuscript. The authors also thank the Editor, Associate Editor, and two anonymous referees for their insightful reviews. The first author was supported by the National Institute

on Aging, Claude D. Pepper Older Americans Independence Centers, Grant P30 AG021334, and NIH-NIA Grant R37 AG19905 at the Johns Hopkins University School of Medicine during the conduct of this research.

References

Barzilai, J. & Borwein, J. M. (1988). Two-point step size gradient methods. *IMA J. Numer. Anal.* **8**, 141–148.

Booth, J. G. & Hobert, J. P. (1999). Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **61**, 265–285.

Caffo, B., Jank, W. & Jones, G. L. (2005). Ascent-based Monte-Carlo expectation-maximization. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **67**, 235–251.

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

Dennis, J. E. Jr & Schnabel, R. B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice Hall, Englewoods Cliffs, NJ.

Frangakis, C. E. & Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics* **58**, 21–29.

Frangakis, C. E., Brookmeyer, R. S., Varadhan, R., Safaeian, M., Vlahov, D. & Strathdee, S. A. (2004). Methodology for evaluating a partially controlled longitudinal treatment using principal stratification, with application to a needle exchange program. *J. Amer. Statist. Assoc.* **99**, 239–249.

Frangakis, C. E. & Varadhan, R. (2004). Systematizing the evaluation of partially controlled studies using principal stratification: from theory to practice. *Statist. Sinica* **14**, 945–947.

Hasselblad, V. (1969). Estimation of finite mixtures of distributions from the exponential family. *J. Amer. Statist. Assoc.* **64**, 1459–1471.

Henrici, P. (1964). *Elements of numerical analysis*. John Wiley, New York.

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

Jamshidian, M. & Jennrich, R. I. (1997). Acceleration of the EM algorithm by using quasi-Newton methods. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **59**, 569–587.

Lancaster, P. (1969). *Theory of matrices*. Academic Press, New York.

Lange, K. (1995). A quasi-Newton acceleration of the EM algorithm. *Statist. Sinica* **5**, 1–18.

Lange, K., Hunter, D. R. & Yang, I. (2000). Optimization transfer using surrogate objective functions. *J. Comput. Graphical Statist.* **9**, 1–20.

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

Liu, C., Rubin, D. B. & Wu, Y. N. (1998). Parameter expansion to accelerate the EM: the PX-EM algorithm. *Biometrika* **85**, 755–770.

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

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

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

Meng, X. L. & Rubin, D. B. (1993). A maximum likelihood estimation via the EM algorithm: a general framework. *Biometrika* **80**, 267–278.

Meng, X. L. & Van Dyk, D. (1997). The EM algorithm – an old folk-song sung to a fast new tune. *J. R. Statist. Soc. Ser. B Statist. Methodol.* **59**, 511–567.

Nievergelt, Y. (1991). Aitken's and Steffensen's accelerations in several variables. *Numer. Math.* **59**, 295–310.

Ortega, J. M. & Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*. Academic Press, New York.

Ortega, J. M. (1973). Stability of difference equations and convergence of iterative processes. *SIAM J. Numer. Anal.* **10**, 268–282.

R Development Core Team (2006). R: A language and environment for statistical computing, Version 2.3.1. Available at: http://www.R-project.org.

Raydan, M. & Svaiter, B. F. (2002). Relaxed steepest descent and Cauchy–Barzilai–Borwein method. *Comput. Optim. Appl.* **21**, 155–167.

Roland, C. & Varadhan, R. (2005). New iterative schemes for nonlinear fixed point problems, with applications to problems with bifurcations and incomplete-data problems. *Appl. Numer. Math.* **55**, 215–226.
Roland, C., Varadhan, R. & Frangakis, C. E. (2007). Squared polynomial extrapolation methods with cycling: an application to the positron emission tomography problem. *Numer. Algorithms* **44**, 159–172.
Smith, D. A., Ford, W. F. & Sidi, A. (1987). Extrapolation methods for vector sequences. *SIAM Rev.* **29**, 199–233.
Varadhan, R. & Roland, C. (2004). Squared extrapolation methods (SQUAREM): a new class of simple and efficient numerical schemes for accelerating the convergence of the EM algorithm, Department of Biostatistics Working Paper, Johns Hopkins University. Available at: http://www.bepress.com/jhubiostat/paper63.
Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Ann. Statist.* **11**, 95–103.

Received March 2007, in final form October 2007

Ravi Varadhan, The Center on Aging and Health, Johns Hopkins University, 2024 E. Monument Street, Suite 2-700, Baltimore, MD 21205, USA.
E-mail: rvaradhan@jhmi.edu