Computational Statistics manuscript No.
(will be inserted by the editor)

Acceleration of the EM algorithm using the vector epsilon algorithm

Mingfeng Wang · Masahiro Kuroda ·
Michio Sakakihara · Zhi Geng

Received: date / Revised version: date

Abstract The Expectation-Maximization (EM) algorithm is a very general and popular iterative computational algorithm to find maximum likelihood estimates from incomplete data and broadly used to statistical analysis with missing data, because of its stability, flexibility and simplicity. However, it is often criticized that the convergence of the EM algorithm is slow. The various algorithms to accelerate the convergence of the EM algorithm have been proposed. The vector ε algorithm of Wynn [14] is used to accelerate the convergence of the EM algorithm in Kuroda and Sakakihara [6]. In this paper, we provide the theoretical evaluation of the convergence of the ε-accelerated EM algorithm. The ε-accelerated EM algorithm does not use the information matrix but only uses the sequence of estimates obtained from iterations of the EM algorithm, and thus it keeps the flexibility and simplicity of the EM algorithm.

Keywords The vector ε algorithm – the EM algorithm – the Louis's EM algorithm – acceleration of convergence

1 Introduction

The Expectation-Maximization (EM) algorithm has been a general and popular algorithm for finding maximum likelihood estimates from incomplete data since Dempster, Laird and Rubin [3]. However, it is often criticized that the convergence of the EM algorithm is slow when the proportion of missing data is high. To improve its convergence, various algorithms incorporating optimization algorithms with faster convergence rate

Mingfeng Wang
Institute of Mathematical Sciences, Peking University, Beijing 100871, CHINA

Masahiro Kuroda
Department of Socio-Information, Okayama University of Science, 1-1 Ridaicho, Okayama 700-0005, JAPAN

Michio Sakakihara
Department of Information Science, Okayama University of Science, 1-1 Ridaicho, Okayama 700-0005, JAPAN

Zhi Geng
Institute of Mathematical Sciences, Peking University, Beijing 100871, CHINA

have been proposed. Louis [9] and Laird, Lange and Stram [7] applied Aitken accelerator to the EM algorithm. Jamshidian and Jennrich [4] used a conjugate gradient accelerator. Lange [8] and Jamshidian and Jennrich [5] proposed a quasi-Newton accelerator. Their techniques require estimates of information matrices or convergence rates during iterations.

Kuroda and Sakakihara [6] proposed the ε-accelerated EM algorithm that accelerates the convergence of the sequence of EM iterations using the vector ε algorithm of Wynn [14]. The merit of the ε-accelerated EM algorithm is that it requires only the sequence of EM iterations but it does not require estimates of information matrices or convergence rates during iterations. In the numerical simulations, Kuroda and Sakakihara [6] demonstrated that the convergence of the ε-accelerated EM algorithm is faster than that of the EM algorithm. But they did not give the theoretical evaluation of the convergence. In this paper, we provide the theoretical evaluation of the convergence of the ε-accelerated EM algorithm.

Section 2 describes the vector ε algorithm. In Section 3, we show the ε-accelerated EM algorithm. In Section 4, we show that the sequence of ε-accelerated EM iterations converges to the stationary point of the sequence of EM iterations and accelerates the convergence of the EM sequence. Section 5 presents numerical experiments to illustrate the behavior of convergence of the ε-accelerated EM algorithm.

2 The vector ε algorithm

The ε algorithm presented by Wynn [13] is utilized to accelerate the convergence of a slowly convergent scalar sequence. The algorithm was also extended to vector sequences by Wynn [14]. It is known that the algorithm is very effective for linearly converging sequences. In this section, we introduce the vector ε algorithm, which is used in the next section to accelerate the EM algorithm.

Let $\theta^{(t)}$ denote a vector of dimension d which converges to a vector θ^* as $t \to \infty$. Let the inverse $[x]^{-1}$ of a vector x be defined by

$$[x]^{-1} = \frac{x}{\|x\|^2}, \qquad \|x\|^2 =< x, x >$$

where $< x, x >$ is the scalar product of x by itself. In general, the vector ε algorithm for a sequence $\{\theta^{(t)}\}_{t \geq 0}$ starts with

$$\varepsilon^{(t,-1)} = 0, \qquad \varepsilon^{(t,0)} = \theta^{(t)},$$

and then generates a vector $\varepsilon^{(t,k+1)}$ by

$$\varepsilon^{(t,k+1)} = \varepsilon^{(t+1,k-1)} + \left[\varepsilon^{(t+1,k)} - \varepsilon^{(t,k)}\right]^{-1}, \qquad k = 0, 1, 2, \ldots. \tag{1}$$

For the case of $k + 1 = 2r + 2$, we have the iteration formula

$$\varepsilon^{(t,2r+2)} = \varepsilon^{(t+1,2r)} + \left[\left[\varepsilon^{(t,2r)} - \varepsilon^{(t+1,2r)}\right]^{-1} + \left[\varepsilon^{(t+2,2r)} - \varepsilon^{(t+1,2r)}\right]^{-1} - \left[\varepsilon^{(t+2,2r-2)} - \varepsilon^{(t+1,2r)}\right]^{-1}\right]^{-1} \tag{2}$$

from equation (1), see Brezinski and Zaglia [2]. For practical implementation, we apply the case of $r = 0$ to equation (2):

$$\varepsilon^{(t,2)} = \varepsilon^{(t+1,0)} + \left[\left[\varepsilon^{(t,0)} - \varepsilon^{(t+1,0)}\right]^{-1} + \left[\varepsilon^{(t+2,0)} - \varepsilon^{(t+1,0)}\right]^{-1} - \left[\varepsilon^{(t+2,-2)} - \varepsilon^{(t+1,0)}\right]^{-1} \right]^{-1}. \tag{3}$$

Then, from initial conditions $\varepsilon^{(t,0)} = \theta^{(t)}$ and $\varepsilon^{(t,-2)} = \infty$ of Brezinski and Zaglia [2], the iteration (3) becomes

$$\varepsilon^{(t,2)} = \theta^{(t+1)} + \left[\left[\theta^{(t)} - \theta^{(t+1)}\right]^{-1} + \left[\theta^{(t+2)} - \theta^{(t+1)}\right]^{-1} \right]^{-1} \tag{4}$$

because $\left[\varepsilon^{(t+2,-2)} - \varepsilon^{(t+1,0)}\right]^{-1} = \left[\infty - \theta^{(t+1,0)}\right]^{-1} = 0$ from the definition of $[x]^{-1}$.

Note that, at each iteration, the computational complexity of the vector ε algorithm is $O(d)$ while the Newton-Raphson algorithm is $O(d^3)$.

3 The ε-accelerated EM algorithm

Let y be the incompletely observed data in a sample space Ω_Y and x be the complete data augmented from y in a sample space Ω_X. Assume that there exists some function $h(x) = y$ relating x to y. Let $f(\cdot|\theta)$ denote a density function with an unknown parameter vector θ in a parameter space Θ. Denote the conditional expectation of the log-likelihood function $\log f(x|\theta)$ given y and θ' by

$$Q(\theta|\theta') = \mathrm{E}[\log f(X|\theta)|y, \theta'].$$

The EM algorithm finds

$$\theta^{(t+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(t)}),$$

at each iteration $t = 0, 1, \ldots$. Given an initial value $\theta^{(0)} \in \Theta$, the ε-accelerated EM algorithm with an additional step of the ε-acceleration is described as follows:

E-step: Find the expectation

$$Q(\theta|\theta^{(t)}) = \mathrm{E}[\log f(X|\theta)|y, \theta^{(t)}].$$

M-step: Find

$$\theta^{(t+1)} = \arg\max_{\theta\in\Theta} Q(\theta|\theta^{(t)}).$$

Then, the M-step produces the sequence of iterates $\{\theta^{(1)}, \ldots, \theta^{(t+1)}\}$.

ε-acceleration: Using $(\theta^{(t+1)}, \theta^{(t)}, \theta^{(t-1)})$, generate the accelerated sequence from

$$\dot{\theta}^{(t-1)} = \theta^{(t)} + \left[\left[\theta^{(t-1)} - \theta^{(t)}\right]^{-1} + \left[\theta^{(t+1)} - \theta^{(t)}\right]^{-1}\right]^{-1}. \qquad (5)$$

Repeat the above steps until

$$||\dot{\theta}^{(t-1)} - \dot{\theta}^{(t-2)}||^2 \leq \delta,$$

where δ is a desired accuracy.

The ε-acceleration step is added to accelerate the convergence of the sequence using the vector ε accelerator and does not depend on statistical models. Thus the ε-accelerated EM algorithm accelerates the convergence without affecting its simplicity and stability. By contrast, the Newton-type accelerations for the EM algorithm are needed to derive the acceleration formula for every statistical model and then calculate the information matrix.

4 Convergence of the ε-accelerated EM algorithm

The convergence of the ε-accelerated EM algorithm needs that the sequence from the EM algorithm converges to a stationary point.

For the EM algorithm, define a mapping $\theta \mapsto M(\theta)$ from Θ to Θ as

$$\theta^{(t+1)} = M(\theta^{(t)}). \qquad (6)$$

Suppose that $\theta^{(t)}$ converges to a stationary point θ^* and derivatives of $M(\theta)$ up to order $r > 2$ exist and are continuous in an open neighborhood of θ^*. Using the Taylor series expansion of equation (6) at θ^*, Meng and Rubin [11] gave the follow equation, for a sufficient large t,

$$\theta^{(t+1)} - \theta^* = DM(\theta^*)(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \qquad (7)$$

where

$$DM(\theta) = \left(\frac{\partial M_j(\theta)}{\partial \theta_i}\right)$$

is the Jacobian matrix for the mapping $M(\theta) = (M_1(\theta), \ldots, M_d(\theta))$. For proof of equation (7), see Appendix A. When t tends to be large, the distance between $\theta^{(t)}$ and θ^* tends to be small, and then equation (7) becomes, for a sufficient large t,

$$\theta^{(t+1)} - \theta^* = \lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \qquad (8)$$

where λ is the largest eigenvalue of $DM(\theta^*)$ (Schafer [12]).

Lemma 1 *From equation (8), we have, for $t \to \infty$,*

$$||\theta^{(t+1)} - \theta^*||^2 = \lambda^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \tag{9}$$

$$||\theta^{(t)} - \theta^{(t+1)}||^2 = (1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \tag{10}$$

$$\begin{aligned}&\langle \theta^{(t)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta^{(t+2)} \rangle = \\ &\qquad \lambda(1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2),\end{aligned} \tag{11}$$

$$\langle \theta^{(t)} - \theta^*, \theta^{(t+1)} - \theta^* \rangle = \lambda ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \tag{12}$$

$$\langle \theta^{(t+1)} - \theta^*, \theta^{(t+2)} - \theta^* \rangle = \lambda^3 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \tag{13}$$

$$\langle \theta^{(t)} - \theta^*, \theta^{(t+2)} - \theta^* \rangle = \lambda^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \tag{14}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product.

Proof See Appendix B. □

Lemma 2 *Let $a = (a_1, \dots, a_d)$ and $b = (b_1, \dots, b_d)$ be two d-dimensional vectors. Then we have*

$$[a^{-1} - b^{-1}]^{-1} = \frac{||b||^2 a - ||a||^2 b}{||a||^2 + ||b||^2 - 2\langle a, b \rangle}. \tag{15}$$

Proof See Appendix C. □

Below we show the convergence of the ε-accelerated EM algorithm.

Theorem 1 *Suppose that the sequence of EM iterates $\{\theta^{(t)}\}_{t \geq 0}$ converges to a stationary point θ^*. The sequence $\{\dot{\theta}^{(t)}\}_{t \geq 0}$ generated by the ε-accelerated EM algorithm converges to a stationary point θ^* of the EM algorithm.*

Proof Let $\eta^{(t)} = [\theta^{(t)} - \theta^{(t+1)}]^{-1} + [\theta^{(t+2)} - \theta^{(t+1)}]^{-1}$. Then we have

$$\dot{\theta}^{(t)} = \theta^{(t+1)} + \Big[[\theta^{(t)} - \theta^{(t+1)}]^{-1} + [\theta^{(t+2)} - \theta^{(t+1)}]^{-1}\Big]^{-1} = \theta^{(t+1)} + [\eta^{(t)}]^{-1}.$$

Since $\lim_{t \to \infty} \theta^{(t)} = \theta^*$, we only need to prove $\lim_{t \to \infty} [\eta^{(t)}]^{-1} = 0$. From equation (15), we have

$$\begin{aligned}&[\eta^{(t)}]^{-1} \\ &= \frac{||\theta^{(t+1)} - \theta^{(t+2)}||^2 (\theta^{(t)} - \theta^{(t+1)}) - ||\theta^{(t)} - \theta^{(t+1)}||^2 (\theta^{(t+1)} - \theta^{(t+2)})}{||\theta^{(t)} - \theta^{(t+1)}||^2 + ||\theta^{(t+1)} - \theta^{(t+2)}||^2 - 2\langle \theta^{(t)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta^{(t+2)} \rangle}.\end{aligned} \tag{16}$$

By Lemma 1, the denominator and numerator of equation (16) can be rewritten as

$$\begin{aligned}&\text{the denominator} \\ &= (1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2) + (1-\lambda)^2 ||\theta^{(t+1)} - \theta^*||^2 \\ &\qquad + o(||\theta^{(t)} - \theta^*||^2) - \{2\lambda(1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2)\} \\ &= (1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2 + (1-\lambda)^2 \{\lambda^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2)\} \\ &\qquad + o(||\theta^{(t)} - \theta^*||^2) - 2\lambda(1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2\end{aligned}$$

$$
\begin{aligned}
&= \{(1-\lambda)^2 + (1-\lambda)^2\lambda^2 - 2\lambda(1-\lambda)^2\}||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2) \\
&= (1-\lambda)^4||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty
\end{aligned}
$$

the numerator

$$
\begin{aligned}
&= \{(1-\lambda)^2||\theta^{(t+1)} - \theta^*||^2 + o(||\theta^{(t+1)} - \theta^*||^2)\}(\theta^{(t)} - \theta^{(t+1)}) \\
&\quad - \{(1-\lambda)^2||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2)\}(\theta^{(t+1)} - \theta^{(t+2)}) \\
&= \{(1-\lambda)^2\lambda^2||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2)\}(\theta^{(t)} - \theta^{(t+1)}) \\
&\quad - \{(1-\lambda)^2||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2)\}(\theta^{(t+1)} - \theta^{(t+2)}), \quad t \to \infty.
\end{aligned}
$$

Since the sequence $\{\theta^{(t)}\}_{t\geq 0}$ converges to a stationary point θ^*, that is, $\lim_{t\to\infty} \theta^{(t)} = \theta^*$, we have

$$\lim_{t\to\infty}(\theta^{(t)} - \theta^{(t+1)}) = \lim_{t\to\infty}(\theta^{(t+1)} - \theta^{(t+2)}) = 0,$$

and thus

$$
\begin{aligned}
&\lim_{t\to\infty}[\eta^{(t)}]^{-1} \\
&= \lim_{t\to\infty} \frac{\{(1-\lambda)^2\lambda^2 + o(1)\}(\theta^{(t)} - \theta^{(t+1)}) - \{(1-\lambda)^2 + o(1)\}(\theta^{(t+1)} - \theta^{(t+2)})}{(1-\lambda)^4 + o(1)} \\
&= 0.
\end{aligned}
$$

Therefore the sequence $\{\dot{\theta}^{(t)}\}_{t\geq 0}$ generated by equation (5) converges to θ^*, because

$$\lim_{t\to\infty}\dot{\theta}^{(t)} = \lim_{t\to\infty}(\dot{\theta}^{(t)} + [\eta^{(t)}]^{-1}) = \lim_{t\to\infty}\theta^{(t)} = \theta^*. \qquad \square$$

Next we evaluate the speed of convergence of the ε-accelerated EM algorithm. For the parameter vector θ, the iterative procedure $\{\theta^{(t)}\}_{t\geq 0}$ is said to converge linearly if

$$c = \lim_{t\to\infty}\frac{||\theta^{(t+1)} - \theta^*||}{||\theta^{(t)} - \theta^*||},$$

where c is some constant and $0 < c < 1$. The sequence of EM iterations converges linearly and the largest eigenvalue λ of $DM(\theta^*)$ corresponds to c. To compare the convergence of the ε-accelerated EM algorithm with that of the EM algorithm, we use the following notion given by Avram [1].

Definition 1 Let A_n be a sequence of scalars, and $\hat{A}_n$ be the sequence generated by applying an extrapolation method ExtM to A_n, where $\hat{A}_n$ is determined from A_m, $0 \leq m \leq L_n$, for some integer L_n, $n = 0, 1, \ldots$. Assume that $\lim_{n\to\infty} A_n = \lim_{n\to\infty}\hat{A}_n = A$. Then we say that $\hat{A}_n$ *converges more quickly* than A_n if

$$\lim_{n\to\infty}\frac{|\hat{A}_n - A|}{|A_{L_n} - A|} = 0.$$

If the above limitation holds, we also say that the extrapolation method ExtM accelerates the convergence of A_n. When A_n is a sequence of vectors in some general vectors space, the definition is still valid, provided we replace $|\hat{A}_n - A|$ and $|A_{L_n} - A|$ everywhere by $\|\hat{A}_n - A\|$ and $\|A_{L_n} - A\|$, respectively, where $\|\cdot\|$ is the norm in the vector space under consideration. The following theorem shows that the convergence of the ε-accelerated EM algorithm is faster than that of the EM algorithm.

Theorem 2 *Assume that $\{\theta^{(t)}\}_{t\geq 0}$ is the sequence of the EM iterations and the sequence $\{\dot{\theta}^{(t)}\}_{t\geq 0}$ is generated by equation (5). Then we have*

$$\lim_{t\to\infty} \frac{||\dot{\theta}^{(t)} - \theta^*||}{||\theta^{(t+2)} - \theta^*||} = 0. \tag{17}$$

That is, $\{\dot{\theta}^{(t)}\}_{t\geq 0}$ converges to θ^* more quickly than $\{\theta^{(t)}\}_{t\geq 0}$ does.

Proof By equation (9), we have for $t \to \infty$,

$$\begin{aligned} ||\theta^{(t+2)} - \theta^*||^2 &= \lambda^2||\theta^{(t+1)} - \theta^*||^2 + o(||\theta^{(t+1)} - \theta^*||^2) \\ &= \lambda^4||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2). \end{aligned}$$

To prove equation (17), we show below that $||\dot{\theta}^{(t)} - \theta^*||^2 = o(||\theta^{(t)} - \theta^*||^2)$ as $t \to \infty$. Let $A = \lambda^2 + o(1)$, $B = 1 + o(1)$ and $C = (1-\lambda)^2 + o(1)$. We have

$$\begin{aligned} &\dot{\theta}^{(t)} - \theta^* \\ &= \theta^{(t+1)} + [\eta^{(t)}]^{-1} - \theta^* \\ &= \theta^{(t)} + \frac{A(\theta^{(t)} - \theta^{(t+1)}) - B(\theta^{(t+1)} - \theta^{(t+2)})}{C} - \theta^* \\ &= \frac{A(\theta^{(t)} - \theta^*) + B(\theta^{(t+2)} - \theta^*) - (A + B - C)(\theta^{(t+1)} - \theta^*)}{C}. \end{aligned}$$

The following equation is proven in Appendix D: for $t \to \infty$,

$$||A\theta^{(t)} + B\theta^{(t+1)} - (A + B - C)\theta^{(t+1)} - C\theta^*||^2 = o(||\theta^{(t)} - \theta^*||^2). \tag{18}$$

Thus we obtain for $t \to \infty$

$$||\dot{\theta}^{(t)} - \theta^*||^2 = \frac{o(||\theta^{(t)} - \theta^*||^2)}{(1-\lambda)^4 + o(1)} = o(||\theta^{(t)} - \theta^*||^2).$$

So we proved Theorem 2. □

5 Numerical Experiments

In this section, we give numerical examples to illustrate the convergence of the EM algorithm, the ε-accelerated EM algorithm and Louis' acceleration EM (Louis EM) algorithm. The computation is implemented with `R` language.

The Louis EM algorithm is a multivariate version of Aitken's acceleration method, and it generates the sequence of iterations $\{\tilde{\theta}^{(t)}\}$ with

$$\tilde{\theta}^{(t)} = \theta^{(t-1)} + (I - J^{(t-1)})^{-1}(\theta^{(t)} - \theta^{(t-1)}), \tag{19}$$

where J is the Jacobian defined by equation (6) and I is the $d \times d$ identity matrix. To estimate $(I - J^{(t-1)})^{-1}$, Louis suggests use of the equation

$$(I - J^{(t-1)})^{-1} = \mathrm{E}[\mathcal{I}(\theta^{(t-1)}|x)|y]\mathcal{I}(\theta^{(t-1)}|y)^{-1}, \tag{20}$$

where

$$\mathcal{I}(\theta^{(t-1)}|x) = -\left.\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(x|\theta)\right|_{\theta=\theta^{(t-1)}},$$
$$\mathcal{I}(\theta^{(t-1)}|y) = -\left.\frac{\partial^2}{\partial\theta\partial\theta^T}\log f(y|\theta)\right|_{\theta=\theta^{(t-1)}}.$$

The Louis EM algorithm iterates the following steps:

EM-step: Find $\theta^{(t)}$ with the E- and M-steps of the EM algorithm.
Aitken acceleration: Estimate $(I - J^{(t-1)})^{-1}$ in equation (20) and compute $\tilde{\theta}^{(t)}$ by equation (19). Then use $\tilde{\theta}^{(t)}$ in the next EM-step.

As pointed out by Louis, the Aitken acceleration with equation (19) is useful in a domain near MLEs and it should not be used until some EM iterations have been performed.

Meilijson [10] also pointed out that the behavior of the Louis EM algorithm with equation (20) is essentially equivalent to the Newton-Raphson algorithm in the neighborhood of the MLEs and its speed of convergence is quadratic. Therefore the Louis EM algorithm is the method to speed up convergence faster than other EM accelerations. However, the Louis algorithm loses the simplicity of the EM algorithm since the Aitken acceleration requires the matrix computation, such as the inverse and product in equation (20) at each iteration. Further the Louis EM algorithm requests its user to provide the formula of information matrix, but the ε-accelerated EM algorithm does not.

Example 1 (Contingency tables with partially classified observations) Consider a 2×2 contingency table with completely and partially classified observations. Let X and Y be dichotomous variables and $\theta = \{\theta_{ij}\}_{i,j=1,2}$ be a set of joint probabilities of X and Y. Denote the cross-classified data of X and Y as $n_{XY} = \{n_{XY}(i,j)\}_{i,j=1,2}$, and the partially classified data of X as $n_X = \{n_X(i)\}_{i=1,2}$ and Y as $n_Y = \{n_Y(j)\}_{j=1,2}$. Assume that the datasets have a multinomial distribution with an unknown parameter θ. The datasets are shown in Table 1. For these data patterns, the convergence of the EM algorithm is quite slow, because its convergence is deeply associated with the proportion of missing data.

The EM algorithm iterates the following two steps:

E-step: Calculate

$$n_X(i,j)^{(t+1)} = n_X(i)\frac{\theta_{ij}^{(t)}}{\sum_j \theta_{ij}^{(t)}}, \qquad n_Y(i,j)^{(t+1)} = n_Y(j)\frac{\theta_{ij}^{(t)}}{\sum_i \theta_{ij}^{(t)}},$$

for all i and j.

M-step: Compute new parameter estimate of θ

$$\theta_{ij}^{(t+1)} = \frac{n_{XY}(i,j) + n_X(i,j)^{(t+1)} + n_Y(i,j)^{(t+1)}}{\sum_{i,j} n_{XY}(i,j) + \sum_i n_X(i) + \sum_j n_Y(j)},$$

for all i and j.

Then, using $(\theta^{(t+1)}, \theta^{(t)}, \theta^{(t-1)})$ of the M-step, the ε-acceleration generates the accelerated sequence

$$\dot{\theta}^{(t-1)} = \theta^{(t)} + \left[\left[\theta^{(t-1)} - \theta^{(t)}\right]^{-1} + \left[\theta^{(t+1)} - \theta^{(t)}\right]^{-1}\right]^{-1}.$$

In Table 2, we give the estimates of θ obtained by the EM, ε-accelerated EM and Louis EM algorithms. Table 3 summarizes the number of iterations for these algorithms for $\delta = 10^{-10}$ and the datasets (a) to (d). For the dataset (a), the ε-accelerated EM algorithm takes 42 iterations using 44 EM iterations to obtain the final values, while the EM algorithm requires 179 iterations. The Louis EM algorithm finds same values after only 7 iterations starting the Aitken acceleration at the third iteration. The third column in Table 4 shows that the ε-accelerated EM algorithm converges 4 to 8 times faster than the EM algorithm and the Louis EM algorithm well accelerates the convergence of the EM algorithm.

The fourth and fifth columns in Table 4 show a comparison based on CPU time. Each CPU time was measured by the function `proc.time`[1]. The fourth column shows CPU time per iteration for each algorithm. For the dataset (a), the CPU time per iteration for the ε-accelerated EM algorithm takes 2.27 times as the EM algorithm, while the Louis EM algorithm requires 8.52 times more CPU time per iteration. Note that the computational cost per iteration for the ε-accelerated EM algorithm is much lower than that of the Louis EM algorithm. The fifth column provides CPU time speedup ratios for these algorithms. For the dataset (a), the ε-accelerated EM algorithm is faster than the EM algorithm by a CPU time speedup ratio of $4.26/2.27 = 1.73$. When applying the Louis EM algorithm, the ratio is 3.00.

Example 2 (Incomplete bivariate normal data) Let (X_1, X_2) be a bivariate normal vector with unknown parameters $\mu = (\mu_1, \mu_2)$ and $\Sigma = (\sigma_{11}, \sigma_{22}, \sigma_{12})$. The incompletely observed data are shown in Table 5, where the observed data from No. 1 to 4 are complete, the data of No. 5, 6, 7 and the data of No. 8, 9, 10 are incomplete with missing values of X_2 and X_1 respectively. The EM algorithm imputes all missing data of incompletely observed data and then estimates μ and Σ using the following iterations:

E-step: Impute the missing value of X_i of the individual with an observed value $X_j = x_j$ by

$$x_i^{(t+1)} = \mu_i^{(t)} + \frac{\sigma_{ij}^{(t)}}{\sigma_{jj}^{(t)}}(x_j - \mu_j^{(t)}).$$

Calculate

$$\begin{aligned} m_1^{(t+1)} &= \sum_{obs.of X_1} x_1 + \sum_{mis.of X_1} x_1^{(t+1)}, \\ m_2^{(t+1)} &= \sum_{obs.of X_2} x_2 + \sum_{mis.of X_2} x_2^{(t+1)}, \\ m_1^{2(t+1)} &= \sum_{obs.of X_1} x_1^2 + \sum_{mis.of X_1} (x_1^{(t+1)})^2 + 10\tau_1^{(t)}, \end{aligned}$$

[1] Times are typically available to 10msec.

$$
\begin{aligned}
m_2^{2(t+1)} &= \sum_{obs.of X_2} x_2^2 + \sum_{mis.of X_2} (x_2^{(t+1)})^2 + 10\tau_2^{(t)}, \\
m_{12}^{(t+1)} &= \sum_{obs.of(X_1,X_2)} x_1 x_2 + \sum_{mis.of X_2} x_1 x_2^{(t+1)} + \sum_{mis.of X_1} x_1^{(t+1)} x_2,
\end{aligned}
$$

where

$$
\tau_1^{(t)} = \sigma_{11}^{(t)} - \frac{(\sigma_{12}^{(t)})^2}{\sigma_{22}^{(t)}}, \quad \tau_2^{(t)} = \sigma_{22}^{(t)} - \frac{(\sigma_{12}^{(t)})^2}{\sigma_{11}^{(t)}}.
$$

M-step: Compute new parameter estimates by

$$
\begin{aligned}
\mu_1^{(t+1)} &= \frac{1}{10} m_1^{(t+1)}, \quad \mu_2^{(t+1)} = \frac{1}{10} m_2^{(t+1)}, \\
\sigma_{11}^{(t+1)} &= \frac{1}{10}\left((m_1^{(t+1)})^2 - m_1^{(t+1)}\right), \quad \sigma_{22}^{(t+1)} = \frac{1}{10}\left((m_2^{(t+1)})^2 - m_2^{(t+1)}\right), \\
\sigma_{12}^{(t+1)} &= \frac{1}{10}\left(m_{12}^{(t+1)} - \frac{1}{10} m_1^{(t+1)} m_2^{(t+1)}\right).
\end{aligned}
$$

Using $(\mu^{(t+1)}, \mu^{(t)}, \mu^{(t-1)})$ and $(\varSigma^{(t+1)}, \varSigma^{(t)}, \varSigma^{(t-1)})$ of the M-step, the ε-acceleration generates the accelerated sequences

$$
\begin{aligned}
\dot{\mu}^{(t-1)} &= \mu^{(t)} + \left[\left[\mu^{(t-1)} - \mu^{(t)}\right]^{-1} + \left[\mu^{(t+1)} - \mu^{(t)}\right]^{-1}\right]^{-1}, \\
\dot{\varSigma}^{(t-1)} &= \varSigma^{(t)} + \left[\left[\varSigma^{(t-1)} - \varSigma^{(t)}\right]^{-1} + \left[\varSigma^{(t+1)} - \varSigma^{(t)}\right]^{-1}\right]^{-1}.
\end{aligned}
$$

The convergences of these algorithms are reported in Table 6. With the desired accuracy $\delta = 10^{-10}$, the MLEs of μ and $\varSigma$ are obtained after 103 iterations for the EM algorithm, but after 51 iterations for the ε-accelerated EM algorithm. Using both algorithms, we can obtain the same MLEs of μ and $\varSigma$:

$$
\mu = (13.673, 13.959), \qquad \varSigma = (53.017, 22.061, 32.910).
$$

Starting the Aitken acceleration at the third, fourth or fifth iteration, the Louis EM algorithm fails to compute $\mathcal{I}(\theta|y)^{-1}$ in equation (20); starting the acceleration at the sixth to 11th iteration, it does not converge to the MLEs. The Louis EM algorithm converges after 23 iterations if the Aitken acceleration starts at the 12th EM iterations. Then the algorithm finds the same MLEs of μ and $\varSigma$ as the EM and ε-accelerated algorithms. Naturally, starting the Aitken acceleration after the sufficient large number of EM iterations, the Louis EM algorithm converges faster to MLEs. If the Aitken acceleration starts at the 18th EM iteration, the Louis EM algorithm converges immediately after only 4 iterations.

Table 7 shows comparisons of the speedup ratios for the ε-accelerated EM algorithm with those of the Louis EM algorithm. For the bivariate normal data, the Louis EM algorithm is more effective in reducing the number of iterations than the ε-accelerated EM algorithm. But CPU time per iteration for the Louis EM algorithm takes 4 times as the ε-accelerated EM algorithm. The ε-accelerated EM algorithm is faster than the EM algorithm by a CPU-time speedup ratio of 1.20, while the Louis EM algorithm decelerates the speed of convergence of the EM algorithm because of the ratio of 0.60.

The Louis EM algorithm is locally quadratic convergence and thus converges theoretically faster than the ε-accelerated EM algorithm whose best speed of convergence is superlinear. But, for the Louis EM algorithm, it may be difficult to determine a starting point. The Louis EM algorithm is also expected numerical instabilities because it requires the computation of the inverse matrix at each iteration. Then the computational cost of the Louis EM algorithm is likely to become more higher as the number of parameters becomes large, while the computational cost of the vector ε algorithm is less expensive. Moreover, as demonstrated in Theorems 1 and 2, the ε-accelerated EM algorithm is guaranteed always to converge faster than the EM algorithm when it converges, but the Louis EM algorithm is not.

Acknowledgements The authors would like to thank the editor and two referees for their valuable comments and helpful suggestions. This research is supported by Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Young Scientists, No 18700279, NSFC and NBRP 2003CB715900.

Appendix A

Proof of equation (7)

For the j-th element of $M(\theta)$, the Taylor series expansion at θ^* is given by

$$\theta_j^{(t+1)} - \theta_j^* = DM_j(\theta^{(t)} - \theta^*) + R_j(\theta^{(t)}, \theta^*),$$

where

$$DM_j(\theta) = \left(\frac{\partial M_j(\theta)}{\partial \theta_1}, \dots, \frac{\partial M_j(\theta)}{\partial \theta_d} \right)$$

and $R_j(\theta^{(t)}, \theta^*)$ is the remainder. For a sufficient large t, the ratio $R_j(\theta^{(t)}, \theta^*)/||\theta^{(t)} - \theta^*||^2$ is bounded, so that $R_j(\theta^{(t)}, \theta^*) = O(||\theta^{(t)} - \theta^*||^2)$. Thus we can obtain equation (7).

Appendix B

Proof of Lemma 1

We show in turn equations (9) to (14).
Proof of equation (9): We have

$$\begin{aligned}
&||\theta^{(t+1)} - \theta^*||^2 \\
&= \langle \theta^{(t+1)} - \theta^*, \theta^{(t+1)} - \theta^* \rangle \\
&= \langle \lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2) \rangle \\
&= \lambda^2 \langle \theta^{(t)} - \theta^*, \theta^{(t)} - \theta^* \rangle + 2\lambda \langle \theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2) \rangle \\
&\quad + \langle O(||\theta^{(t)} - \theta^*||^2), O(||\theta^{(t)} - \theta^*||^2) \rangle, \quad t \to \infty.
\end{aligned}$$

By $\theta^* = \lim_{t\to\infty} \theta^{(t)}$, we get

$$\begin{aligned}
&\lim_{t\to\infty} \frac{\langle \theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2)\rangle}{||\theta^{(t)} - \theta^*||^2} \\
&= \lim_{t\to\infty} \frac{||\theta^{(t)} - \theta^*||^2 \langle \theta^{(t)} - \theta^*, O(1)\rangle}{||\theta^{(t)} - \theta^*||^2} = \lim_{t\to\infty} \langle \theta^{(t)} - \theta^*, O(1)\rangle = 0
\end{aligned}$$

and

$$\begin{aligned}
&\lim_{t\to\infty} \frac{\langle O(||\theta^{(t)} - \theta^*||^2), O(||\theta^{(t)} - \theta^*||^2)\rangle}{||\theta^{(t)} - \theta^*||^2} \\
&= \lim_{t\to\infty} \frac{||\theta^{(t)} - \theta^*||^4 \langle O(1), O(1)\rangle}{||\theta^{(t)} - \theta^*||^2} \\
&= \lim_{t\to\infty} ||\theta^{(t)} - \theta^*||^2 \langle O(1), O(1)\rangle = \lim_{t\to\infty} ||\theta^{(t)} - \theta^*||^2 O(1) = 0.
\end{aligned}$$

Thus we obtain that as $t \to \infty$,

$$\langle \theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2)\rangle = o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty$$

and

$$\langle O(||\theta^{(t)} - \theta^*||^2), O(||\theta^{(t)} - \theta^*||^2)\rangle = o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.$$

Then we have

$$\begin{aligned}
||\theta^{(t+1)} - \theta^*||^2 &= \lambda^2 ||\theta^{(t)} - \theta^*||^2 + 2\lambda o(||\theta^{(t)} - \theta^*||^2) + o(||\theta^{(t)} - \theta^*||^2) \\
&= \lambda^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.
\end{aligned}$$

Proof of equation (10): From

$$\begin{aligned}
\theta^{(t)} - \theta^{(t+1)} &= (\theta^{(t)} - \theta^*) - (\theta^{(t+1)} - \theta^*) \\
&= (\theta^{(t)} - \theta^*) - \lambda(\theta^{(t)} - \theta^*) - O(||\theta^{(t)} - \theta^*||^2) \\
&= (1-\lambda)(\theta^{(t)} - \theta^*) - O(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty
\end{aligned}$$

we have

$$\begin{aligned}
&||\theta^{(t)} - \theta^{(t+1)}||^2 \\
&= \langle \theta^{(t)} - \theta^{(t+1)}, \theta^{(t)} - \theta^{(t+1)}\rangle \\
&= \langle (1-\lambda)(\theta^{(t)} - \theta^*) - O(||\theta^{(t)} - \theta^*||^2), (1-\lambda)(\theta^{(t)} - \theta^*) - O(||\theta^{(t)} - \theta^*||^2)\rangle \\
&= (1-\lambda)^2 \langle \theta^{(t)} - \theta^*, \theta^{(t)} - \theta^*\rangle - 2(1-\lambda)\langle \theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2)\rangle \\
&\quad + \langle O(||\theta^{(t)} - \theta^*||^2), O(||\theta^{(t)} - \theta^*||^2)\rangle, \quad t \to \infty.
\end{aligned}$$

In the same way as the proof of equation (9), we can obtain

$$||\theta^{(t)} - \theta^{(t+1)}||^2 = (1-\lambda)^2 ||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.$$

Proof of equation (11) : From

$$\begin{aligned}\theta^{(t+2)} - \theta^* &= \lambda(\theta^{(t+1)} - \theta^*) + O(||\theta^{(t+1)} - \theta^*||^2)\\ &= \lambda\left(\lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2)\right) + O(||\theta^{(t+1)} - \theta^*||^2)\\ &= \lambda^2(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty\end{aligned}$$

we have

$$\begin{aligned}&\theta^{(t+1)} - \theta^{(t+2)}\\ &= (\theta^{(t+1)} - \theta^*) - (\theta^{(t+2)} - \theta^*)\\ &= (\theta^{(t+1)} - \theta^*) - \left(\lambda(\theta^{(t+1)} - \theta^*) + O(||\theta^{(t+1)} - \theta^*||^2)\right)\\ &= (1-\lambda)\left(\lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2)\right) - O(||\theta^{(t+1)} - \theta^*||^2)\\ &= \lambda(1-\lambda)(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.\end{aligned}$$

Then we get

$$\begin{aligned}&\langle\theta^{(t)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta^{(t+2)}\rangle\\ &= \langle(1-\lambda)(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \lambda(1-\lambda)(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2)\rangle\\ &= \lambda(1-\lambda)^2||\theta^{(t)} - \theta^*||^2 + (1-\lambda)\langle\theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2)\rangle\\ &\quad + \lambda(1-\lambda)\langle\theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2)\rangle + \langle O(||\theta^{(t)} - \theta^*||^2), O(||\theta^{(t)} - \theta^*||^2)\rangle\\ &= \lambda(1-\lambda)^2||\theta^{(t)} - \theta^*||^2 + (1-\lambda)o(||\theta^{(t)} - \theta^*||^2) + \lambda(1-\lambda)o(||\theta^{(t)} - \theta^*||^2)\\ &\quad + O(||\theta^{(t)} - \theta^*||^4), \quad t \to \infty.\end{aligned}$$

Thus we obtain

$$\begin{aligned}&\langle\theta^{(t)} - \theta^{(t+1)}, \theta^{(t+1)} - \theta^{(t+2)}\rangle\\ &= \lambda(1-\lambda)^2||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.\end{aligned}$$

Proof of equation (12): From

$$\theta^{(t+1)} - \theta^* = \lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty$$

we have

$$\begin{aligned}&\langle\theta^{(t)} - \theta^*, \theta^{(t+1)} - \theta^*\rangle\\ &= \langle\theta^{(t)} - \theta^*, \lambda(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2)\rangle\\ &= \lambda||\theta^{(t)} - \theta^*||^2 + \langle\theta^{(t)} - \theta^*, O(||\theta^{(t)} - \theta^*||^2)\rangle\\ &= \lambda||\theta^{(t)} - \theta^*||^2 + o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.\end{aligned}$$

Proof of equation (13): From the proof of equation (11), we have

$$\theta^{(t+2)} - \theta^* = \lambda^2(\theta^{(t)} - \theta^*) + O(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.$$

Thus we get

$$\begin{aligned}
&\langle\theta^{(t+1)}-\theta^*,\theta^{(t+2)}-\theta^*\rangle\\
&=\langle\lambda(\theta^{(t)}-\theta^*)+O(||\theta^{(t)}-\theta^*||^2),\lambda^2(\theta^{(t)}-\theta^*)+O(||\theta^{(t)}-\theta^*||^2)\rangle\\
&=\lambda^3||\theta^{(t)}-\theta^*||^2+o(||\theta^{(t)}-\theta^*||^2),\quad t\to\infty.
\end{aligned}$$

Proof of equation (14): We have

$$\begin{aligned}
&\langle\theta^{(t)}-\theta^*,\theta^{(t+2)}-\theta^*\rangle\\
&=\langle\theta^{(t)}-\theta^*,\lambda^2(\theta^{(t)}-\theta^*)+O(||\theta^{(t)}-\theta^*||^2)\rangle\\
&=\lambda^2\langle\theta^{(t)}-\theta^*,\theta^{(t)}-\theta^*\rangle+\langle\theta^{(t)}-\theta^*,O(||\theta^{(t)}-\theta^*||^2)\rangle\\
&=\lambda^2||\theta^{(t)}-\theta^*||^2+o(||\theta^{(t)}-\theta^*||^2),\quad t\to\infty.
\end{aligned}$$

Appendix C

Proof of Lemma 2

We have

$$\begin{aligned}
[a^{-1}-b^{-1}]&=\left[\frac{a}{||a||^2}-\frac{b}{||b||^2}\right]^{-1}=\left[\frac{||b||^2a-||a||^2b}{||a||^2||b||^2}\right]^{-1}\\
&=\frac{||b||^2a-||a||^2b}{||a||^2||b||^2}\Bigg/\left|\left|\frac{||b||^2a-||a||^2b}{||a||^2||b||^2}\right|\right|^2\\
&=\frac{||a||^2||b||^2(||b||^2a-||a||^2b)}{\sum_i(||b||^2a_i-||a||^2b_i)^2}\\
&=\frac{||a||^2||b||^2(||b||^2a-||a||^2b)}{\sum_i(||b||^4a_i^2+||a||^4b_i^2-2||a||^2||b||^2a_ib_i)}\\
&=\frac{||a||^2||b||^2(||b||^2a-||a||^2b)}{||b||^4||a||^2+||a||^4||b||^2-2||a||^2||b||^2\langle a,b\rangle}\\
&=\frac{||b||^2a-||a||^2b}{||a||^2+||b||^2-2\langle a,b\rangle}.
\end{aligned}$$

Appendix D

Proof of equation (18) in Theorem 2

Let $\delta^{(t)}=\theta^{(t)}-\theta^*$. Then we have

$$\begin{aligned}
&||A(\theta^{(t)}-\theta^*)+B(\theta^{(t+2)}-\theta^*)-(A+B-C)(\theta^{(t+1)}-\theta^*)||^2\\
&=||A\delta^{(t)}+B\delta^{(t+2)}-(A+B-C)\delta^{(t+1)}||^2\\
&=\langle A\delta^{(t)}+B\delta^{(t+2)}-(A+B-C)\delta^{(t+1)},A\delta^{(t)}+B\delta^{(t+2)}-(A+B-C)\delta^{(t+1)}\rangle
\end{aligned}$$

$$
\begin{aligned}
&= \langle A\delta^{(t)} + B\delta^{(t+2)}, A\delta^{(t)} + B\delta^{(t+2)}\rangle + \langle (A+B-C)\delta^{(t+1)}, (A+B-C)\delta^{(t+1)}\rangle \\
&\quad - 2\langle A\delta^{(t)} + B\delta^{(t+2)}, (A+B-C)\delta^{(t+1)}\rangle \\
&= A^2||\delta^{(t)}||^2 + 2AB\langle\delta^{(t)}, \delta^{(t+2)}\rangle + B^2||\delta^{(t+2)}||^2 + (A+B-C)^2||\delta^{(t+1)}||^2 \\
&\quad - 2A(A+B-C)\langle\delta^{(t)}, \delta^{(t+1)}\rangle - 2B(A+B-C)\langle\delta^{(t+2)}, \delta^{(t+1)}\rangle \\
&= A^2||\delta^{(t)}||^2 + 2AB\lambda^2||\delta^{(t)}||^2 + B^2\lambda^4||\delta^{(t)}||^2 + (A+B-C)^2\lambda^2||\delta^{(t)}||^2 \\
&\quad - 2A(A+B-C)\lambda||\delta^{(t)}||^2 - 2B(A+B-C)\lambda^3||\delta^{(t)}||^2 + o(||\delta^{(t)}||^2).
\end{aligned}
$$

Substituting $A = \lambda^2 + o(1)$, $B = 1 + o(1)$ and $C = (1-\lambda)^2 + o(1)$ into the above equation, we can obtain

$$
\begin{aligned}
&A^2 + 2AB\lambda^2 + B^2\lambda^4 + (A+B-C)^2\lambda^2 - 2A(A+B-C)\lambda - 2B(A+B-C)\lambda^3 \\
&\quad = o(1).
\end{aligned}
$$

Thus we showed

$$
\begin{aligned}
&||A(\theta^{(t)} - \theta^*) + B(\theta^{(t+2)} - \theta^*) - (A+B-C)(\theta^{(t+1)} - \theta^*)||^2 \\
&= o(||\theta^{(t)} - \theta^*||^2), \quad t \to \infty.
\end{aligned}
$$

References

1. Avram S (2003) Practical Extrapolation Methods, Theory and Applications. Cambridge University Press, Cambridge
2. Brezinski C, Zaglia MR (1991) Extrapolation methods: theory and practice. Elsevier Science Ltd
3. Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B. 39:1-22.
4. Jamshidian M, Jennrich RI (1993) Conjugate gradient acceleration of the EM algorithm. J. Amer. Statist. Assoc. 88:221-228.
5. Jamshidian M, Jennrich RI (1997) Acceleration of the EM algorithm by using quasi-Newton methods. J. R. Stat. Soc. Ser. B. 59:569-587.
6. Kuroda M, Sakakihara M (2006) Accelerating the convergence of the EM algorithm using the vector ε algorithm. Comput. Statist. Data Anal. 51:1549-1561
7. Laird NM, Lange K, Stram DO (1987) Maximum likelihood computations with repeated measures: application of the EM algorithm. J. Am. Statist. Ass. 82:97-105.
8. Lange K. (1995) A quasi Newton acceleration of the EM algorithm. Statist. Sin. 5:1-18.
9. Louis TA (1982) Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B. 44:226-233.
10. Meilijson I (1989) A fast improvement to the EM algorithm on its own terms. J. Roy. Statist. Soc. Ser. B 51:127-138.
11. Meng XL, Rubin DB (1994) On the global and componentwise rates of convergence of the EM algorithm. Linear Algebra Appli. 199:413-425.
12. Schafer JL (1997) Analysis of Incomplete Multivariate Data. Chapman & Hall/CRC, London.
13. Wynn P (1961) The epsilon algorithm and operational formulas of numerical analysis. Math. Comp. 15:151-158.
14. Wynn P. (1962) Acceleration techniques for iterated vector and matrix problems. Math. Comp., 16:301-322.

Table 1 Contingency table with completely and partially classified data

	n_Y		n_X		n_{XY}			
			$i=1$	$i=2$	$i=1$		$i=2$	
	$j=1$	$j=2$			$j=1$	$j=2$	$j=1$	$j=2$
(a)	100	60	300	200	5	4	2	1
(b)	250	150	300	200	5	4	2	1
(c)	500	300	300	200	5	4	2	1
(d)	1000	600	300	200	5	4	2	1

Table 2 Estimates of $\theta = \{\theta_{ij}\}_{i,j=1,2}$

	θ_{11}	θ_{12}	θ_{21}	θ_{22}
(a)	0.3465	0.2570	0.2769	0.1197
(b)	0.3469	0.2565	0.2774	0.1192
(c)	0.3471	0.2564	0.2776	0.1190
(d)	0.3472	0.2563	0.2776	0.1189

Table 3 The number of iterations for $\delta = 10^{-10}$

	EM	ε-accelerated EM	Louis EM
(a)	179	42	7
(b)	225	27	7
(c)	277	37	7
(d)	335	61	7

Table 4 Speedup ratios for each contingency table

		Iteration speedup ratio	CPU-time ratio	CPU-time speedup ratio
(a)	EM	1.00	1.00	1.00
	ε-accelerated EM	4.26	2.27	1.73
	Louis EM	25.57	8.52	3.00
(b)	EM	1.00	1.00	1.00
	ε-accelerated EM	8.33	3.16	2.64
	Louis EM	32.14	8.04	4.00
(c)	EM	1.00	1.00	1.00
	ε-accelerated EM	7.49	2.93	2.56
	Louis EM	39.57	9.89	4.00
(d)	EM	1.00	1.00	1.00
	ε-accelerated EM	5.49	2.13	2.58
	Louis EM	47.86	9.57	5.00

Table 5 Incomplete bivariate normal data

No.	1	2	3	4	5	6	7	8	9	10
X_1	8	11	16	18	25	9	13	*	*	*
X_2	10	14	16	15	*	*	*	15	20	4

("*" indicates a missing value)

Table 6 The number of iterations of the Louis EM algorithm starting Aitken accelerated from several EM iterations

Algorithm	The number of iterations	Convergence
EM	103	
ε-accelerated EM	51	
Louis EM	— (3)	fail to compute $\mathcal{I}(\theta\|y)^{-1}$
	— (4)	
	— (5)	
	52 (6)	not converge to MLEs or
	⋮	fail to compute $\mathcal{I}(\theta\|y)^{-1}$
	47 (11)	
	23 (12)	converge to MLEs
	⋮	
	22 (18)	

Table 7 Speedup ratios for the incomplete bivariate normal data

	Iteration speedup ratio	CPU-time ratio	CPU-time speedup ratio
EM	1.00	1.00	1.00
ε-accelerated EM	2.02	1.68	1.20
Louis EM	4.47	7.46	0.60