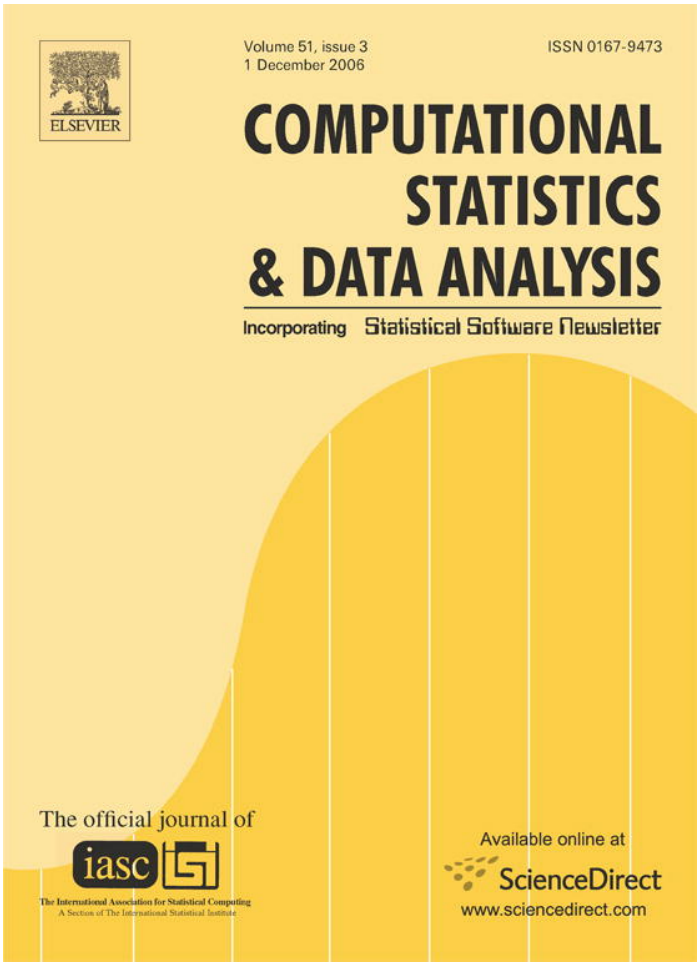

Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics & Data Analysis 51 (2006) 1549–1561
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Accelerating the convergence of the EM algorithm using the vector ε algorithm

Masahiro Kuroda[a,*], Michio Sakakihara[b]

[a] *Department of Socio-Information, Okayama University of Science, 1-1 Ridaicho, Okayama, Japan*
[b] *Department of Information Science, Okayama University of Science, 1-1 Ridaicho, Okayama, Japan*

Received 20 September 2005; received in revised form 17 April 2006; accepted 9 May 2006
Available online 2 June 2006

Abstract

The EM algorithm of Dempster, Laird and Rubin [1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1–22] is a very general and popular iterative computational algorithm that is used to find maximum likelihood estimates from incomplete data and is widely used to perform statistical analysis with missing data, because of its stability, flexibility and simplicity. However, a common criticism is that the convergence of the EM algorithm is slow. Various algorithms to accelerate the convergence of the EM algorithm have been proposed. In this paper, we propose the "ε-accelerated EM algorithm" that speeds up the convergence of the EM sequence via the vector ε algorithm of Wynn [1962. Acceleration techniques for iterated vector and matrix problems. Math. Comp. 16, 301–322]. We also demonstrate its theoretical properties. The ε-accelerated EM algorithm has been successfully extended to the EM algorithm without affecting its stability, flexibility and simplicity. Numerical experiments illustrate the potential of the ε-accelerated EM algorithm.

Keywords: Vector ε algorithm; EM algorithm; Acceleration of convergence

1. Introduction

The EM algorithm of Dempster et al. (1977) is a very general and popular iterative computational algorithm that is used to find maximum likelihood estimates (MLEs) from incomplete data and is widely used to perform statistical analysis with missing data, because of its stability, flexibility and simplicity. However, a common criticism is that the convergence of the EM algorithm is slow, since its speed of convergence is linear. In particular, the convergence of the EM algorithm is quite slow when the proportion of missing data is high.

In order to speed up the convergence of the EM algorithm, various acceleration algorithms have been proposed. Louis (1982) suggested the EM algorithm incorporating the multivariate version of Aitken's acceleration method. Jamshidian and Jennrich (1993) proposed an acceleration algorithm based on conjugate gradients. Lange (1995) used a quasi-Newton algorithm to accelerate the EM algorithm. Using the Newton-type accelerators requires the computation of a matrix inversion at each iteration. Then its computation is likely to become rapidly complicated as the number of

* Corresponding author. Tel.: +81 86 2569741; fax: +81 86 2568006.
E-mail address: kuroda@soci.ous.ac.jp (M. Kuroda).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.05.004

parameters increases, and numerical instabilities are also expected. Therefore, these accelerations do not have some of the attractive features of the EM algorithm, such as stability, flexibility and simplicity.

Some extensions of the EM algorithm have been proposed. Meng and Rubin (1993) presented the expectation conditional maximization (ECM) algorithm that improves the maximization process in the M-step. Liu and Rubin (1994) suggested the expectation conditional maximization either (ECME) algorithm that is an extension of the ECM algorithm. McLachlan and Krishnan (1997) provided a comprehensive account of the EM algorithm. Geng et al. (2000) presented the partial imputation EM algorithm that imputes a part of missing data such that the observed data and imputed data form a monotone data pattern.

In this paper, we propose the "ε-accelerated EM algorithm" that accelerates the convergence of the EM sequence using the vector ε algorithm of Wynn (1961). Section 2 describes the vector ε algorithm. In Section 3, we presented the ε-accelerated EM algorithm. Then we give the key results that the sequence generated by the ε-accelerated EM algorithm converges to the stationary point of the EM sequence and, for scalar sequences, it converges faster than the EM sequence. Section 5 examines the performance and properties of the ε-accelerated EM algorithm using numerical experiments.

2. The vector ϵ algorithm

The ε algorithm of Wynn (1961) is utilized to accelerate the convergence of a slowly convergent scalar sequence. The algorithm has also been extended to vector sequences by Wynn (1962). It is known that the algorithm is very effective for linearly converging sequences. In this section, we illustrate the vector ε algorithm to obtain the proposed acceleration method for the EM algorithm.

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

$$[x]^{-1} = \frac{x}{\|x\|^2}, \quad \|x\|^2 = \langle x, x \rangle$$

in which $\langle x, x \rangle$ is the scalar product of x by itself.

In general, the vector ε algorithm for a sequence $\left\{\theta^{(t)}\right\}_{t \geqslant 0}$ starts with

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

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

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

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

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

from Eq. (1), see Brezinski and Zaglia (1991). For practical implementation, we apply the case of $r = 0$ to Eq. (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 the initial conditions $\varepsilon^{(t,0)} = \theta^{(t)}$ and $\varepsilon^{(t,-2)} = \infty$ of Brezinski and Zaglia (1991), the iteration (3) becomes as follows:

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

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 vector ε algorithm requires only $O(d)$ arithmetic operations while the Newton–Raphson and quasi-Newton algorithms are achieved at $O(d^3)$ and $O(d^2)$ and thus the computational cost is likely to become more expensive as d becomes large.

3. The ϵ-accelerated EM algorithm

Let y be observed data with a sample space Ω_Y and x be complete data augmented by y with a sample space Ω_X. We assume that there exists a many-to-one mapping from Ω_X to Ω_Y. Instead of observing x in Ω_X, we observe $y = y(x)$ in Ω_Y. Let $f(x|\theta)$ denote the probability density function of x and $g(y|\theta)$ denote the probability density function of y, where θ is an unknown parameter vector with a parameter space Θ.

Define the conditional expectation of the log-likelihood function $\log f(x|\theta)$ given y and θ' as

$$Q\left(\theta|\theta'\right) = E\left[\log f(X|\theta)|y, \theta'\right].$$

The EM algorithm chooses

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

at each iteration $t = 1, 2, \ldots$.

Given an initial value $\theta^{(0)} \in \Theta$, the ε-accelerated EM algorithm incorporating the ε-acceleration process performs the following steps:

E-step: Calculate

$$Q\left(\theta|\theta^{(t-1)}\right) = E\left[\log f(X|\theta)|y, \theta^{(t-1)}\right].$$

M-step: Choose $\theta^{(t)}$ such that

$$Q\left(\theta^{(t)}|\theta^{(t-1)}\right) \geqslant Q\left(\theta|\theta^{(t-1)}\right),$$

for all $\theta \in \Theta$.

ε-acceleration: Calculate

$$\dot{\theta}^{(t-2)} = \theta^{(t-1)} + \left[\left[\theta^{(t-2)} - \theta^{(t-1)}\right]^{-1} + \left[\theta^{(t)} - \theta^{(t-1)}\right]^{-1}\right]^{-1}$$

from Eq. (4) and check the convergence using

$$\left\|\dot{\theta}^{(t-2)} - \dot{\theta}^{(t-3)}\right\|_\infty \leqslant \delta,$$

where $\|x\|_\infty = \max_i \{|x_i|\}$ for a vector $x = \{x_i\}$ and δ is a desired accuracy.

Note that the ε-accelerated EM algorithm does not improve the E- and M-steps but it accelerates the convergence of the EM sequence using the ε-acceleration process.

4. Convergence of the ϵ-accelerated EM algorithm

The EM algorithm is a first-order successive substitution method and implicitly defines a map $\theta \to M(\theta)$ from Θ to Θ such that

$$\theta^{(t+1)} = M\left(\theta^{(t)}\right).$$

Suppose that $\theta^{(t)}$ converges to a stationary point θ^* and that $M(\theta)$ is differentiable at θ^*. Expanding $M\left(\theta^{(t)}\right)$ in a Taylor series about θ^*, Meng and Rubin (1994) gave the following equation:

$$\theta^{(t+1)} - \theta^* = J^*\left(\theta^{(t)} - \theta^*\right) + O\left(\left\|\theta^{(t)} - \theta^*\right\|^2\right), \quad t \to \infty, \tag{5}$$

where J^* is the Jacobian matrix for $M(\theta)$ evaluated at θ^*. For sufficiently large t, the distance between $\theta^{(t)}$ and θ^* tends to be small, and then Eq. (5) becomes

$$\theta^{(t+1)} - \theta^* = \lambda\left(\theta^{(t)} - \theta^*\right) + \mathrm{O}\left(\left\|\theta^{(t)} - \theta^*\right\|^2\right), \quad t \to \infty, \tag{6}$$

where λ is the largest eigenvalue of J^* (Schafer, 1997).

For the parameter vector θ, the iterative sequence $\left\{\theta^{(t)}\right\}_{t\geqslant 0}$ is said to converge linearly if

$$c = \lim_{t\to\infty} \frac{\left\|\theta^{(t+1)} - \theta^*\right\|}{\left\|\theta^{(t)} - \theta^*\right\|}, \tag{7}$$

where c is some constant, $0 < c < 1$. The EM sequence converges linearly and λ corresponds to c.

Now we can show the convergence of the ε-acceleration process.

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

Proof. See Appendix B.

Next we investigate the speed of convergence of the ε-accelerated EM algorithm for a scalar sequence $\left\{\theta^{(t)}\right\}_{t\geqslant 0}$. To compare the speed of convergence of the ε-accelerated EM algorithm with that of the EM algorithm, we provide the following notion (Brezinski and Zaglia, 1991).

Definition 2. Let $\left\{\hat{\theta}^{(t)}\right\}_{t\geqslant 0}$ be a scalar sequence obtained by applying an extrapolation method to $\left\{\theta^{(t)}\right\}_{t\geqslant 0}$. Assume that $\lim_{t\to\infty}\theta^{(t)} = \lim_{t\to\infty}\hat{\theta}^{(t)} = \theta^*$. If

$$\lim_{t\to\infty} \frac{\left|\hat{\theta}^{(t)} - \theta^*\right|}{\left|\theta^{(t)} - \theta^*\right|} = 0,$$

then we say that the sequence $\left\{\hat{\theta}^{(t)}\right\}_{t\geqslant 0}$ converges to θ^* faster than $\left\{\theta^{(t)}\right\}_{t\geqslant 0}$ or the extrapolation method accelerates the convergence of $\left\{\theta^{(t)}\right\}_{t\geqslant 0}$.

For a scalar sequence $\left\{\theta^{(t)}\right\}_{t\geqslant 0}$, Eq. (4) is formulated by

$$\begin{aligned}\dot{\theta}^{(t)} &= \theta^{(t+1)} + \left(\frac{1}{\theta^{(t)} - \theta^{(t+1)}} + \frac{1}{\theta^{(t+2)} - \theta^{(t+1)}}\right)^{-1} \\ &= \theta^{(t+1)} + \frac{\left(\theta^{(t)} - \theta^{(t+1)}\right)\left(\theta^{(t+2)} - \theta^{(t+1)}\right)}{\theta^{(t+2)} - 2\theta^{(t+1)} + \theta^{(t)}}\end{aligned} \tag{8}$$

and is identical to the Aitken δ^2 method of Aitken (1926). Traub (1964) proved that the Aitken δ^2 method accelerates the convergence of linear convergent sequences in the sense that

$$\lim_{t\to\infty} \frac{\left|\hat{\theta}^{(t)} - \theta^*\right|}{\left|\theta^{(t+2)} - \theta^*\right|} = 0.$$

From the fact that the EM sequences converge linearly, we can see that the ε-accelerated EM algorithm accelerates the convergence of the EM sequence.

Brezinski and Zaglia (1991) have completely described the convergence and acceleration properties of the vector ε algorithm, only for special classes of vector sequences. Thus, the mathematical consideration of the ε-accelerated EM algorithm is not trivial.

Louis (1982) gave an acceleration formula that resembles Eq. (4) based on the multivariate version of the Aitken's acceleration method as follows:

$$\tilde{\theta}^{(t)} = \theta^{(t-1)} + \left(I_d - J^{(t-1)}\right)^{-1} \left(\theta^{(t)} - \theta^{(t-1)}\right), \tag{9}$$

where I_d denotes the $d \times d$ identity matrix and $J^{(t-1)}$ is the Jacobian matrix for $M(\theta)$ evaluated at $\theta^{(t-1)}$. In order to estimate $\left(I_d - J^{(t-1)}\right)^{-1}$, Louis suggested making use of the equation

$$\left(I_d - J^{(t-1)}\right)^{-1} = E\left[\mathscr{I}\left(\theta^{(t-1)}|X\right)|y\right] I\left(\theta^{(t-1)}|y\right)^{-1}, \tag{10}$$

where

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

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

After performing the E- and M-steps, Aitken's acceleration using Eq. (9) generates $\tilde{\theta}^{(t)}$ and then uses it in the next E- and M-steps. Meilijson (1989) pointed out that Louis's acceleration method using Eq. (9) is essentially equivalent to the Newton–Raphson algorithm in the neighborhood of MLEs.

The ε-accelerated EM algorithm always uses acceleration (4), which does not depend on statistical models. By contrast, the Newton-type accelerations for the EM algorithm are needed to derive the acceleration formula for every statistical model. Thus, the ε-accelerated EM algorithm accelerates the convergence without affecting its simplicity and stability.

5. Numerical experiments

In this section, we provide numerical experiments to investigate how much faster the ε-accelerated EM algorithm converges than the EM algorithm. Note that the speed of convergence of the ε-accelerated EM algorithm does not depend on statistical models but on the EM sequence. To examine the performance of the ε-accelerated EM algorithm, we apply the algorithm to data sets for which the speed of convergence of the EM algorithm is quite slow.

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 $\theta = \left\{p_{ij}\right\}_{i,j=1,2}$ 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 data sets have a multinomial distribution with an unknown parameter θ.

The data sets 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. In Table 2, we summarize the number of iterations for the EM and ε-accelerated EM algorithms for each $\delta = 10^{-5}$–10^{-8} and the data sets (a)–(e), and plot these results in Fig. 1. Table 3 gives the estimates of θ by the EM and ε-accelerated EM algorithms.

As shown in Table 2 and Fig. 1, the EM algorithm increases linearly with the number of iterations as the data set changes from (a) to (e), while there is little variation in the number of iterations for the ε-accelerated EM algorithm and its convergence is significantly faster. For example, for $\delta = 10^{-6}$ and the data set (d), the ε-accelerated EM algorithm takes only 41 iterations using 43 EM iterations to obtain the final values, while the EM algorithm takes 476 iterations

Table 1
Contingency table with completely and partially classified data

	n_Y		n_X		n_{XY}			
	$j=1$	$j=2$	$i=1$	$i=2$	$i=1$		$i=2$	
					$j=1$	$j=2$	$j=1$	$j=2$
(a)	50	30	300	200	5	4	2	1
(b)	100	60	300	200	5	4	2	1
(c)	250	150	300	200	5	4	2	1
(d)	500	300	300	200	5	4	2	1
(e)	1000	600	300	200	5	4	2	1

Table 2
The number of iterations for each δ

δ		(a)	(b)	(c)	(d)	(e)
10^{-5}	EM	123	136	166	192	198
	ε-accelerated EM	58	40	27	36	59
10^{-6}	EM	253	282	364	476	656
	ε-accelerated EM	72	48	32	41	68
10^{-7}	EM	383	429	561	761	1114
	ε-accelerated EM	84	64	79	90	86
10^{-8}	EM	513	575	759	1045	1572
	ε-accelerated EM	119	136	179	234	313

to achieve the same values. For other data sets, the EM algorithm requires the number of iterations roughly 3–10 times greater than that of the ε-accelerated EM algorithm. Table 3 demonstrates that the estimates by the EM algorithm for $\delta = 10^{-5}$ match only three digits to the MLEs that are the estimates with $\delta = 10^{-7}$ and 10^{-8}. Then the ε-accelerated EM algorithm finds estimates that are identical to the MLEs.

Next we investigate the speed of convergence of the ε-accelerated EM algorithm. In these numerical experiments, the (i, j)th component-wise speeds of convergence of the EM and ε-accelerated EM algorithms are assessed as

$$R_{ij} = \lim_{t\to\infty} r_{ij}^{(t)} = \lim_{t\to\infty} \frac{\left| p_{ij}^{(t)} - p_{ij}^{\mathrm{MLE}} \right|}{\left| p_{ij}^{(t-1)} - p_{ij}^{\mathrm{MLE}} \right|},$$

$$\dot{R}_{ij} = \lim_{t\to\infty} \dot{r}_{ij}^{(t)} = \lim_{t\to\infty} \frac{\left| \dot{p}_{ij}^{(t)} - p_{ij}^{\mathrm{MLE}} \right|}{\left| p_{ij}^{(t+2)} - p_{ij}^{\mathrm{MLE}} \right|}.$$

If there exist MLEs of θ, $\theta^{\mathrm{MLE}} = \left\{ p_{ij}^{\mathrm{MLE}} \right\}_{i,j=1,2}$, and the ε-accelerated EM algorithm accelerates the convergence of the sequence $\left\{ \theta^{(t)} \right\}_{t\geqslant 0}$, then it holds that $\dot{R}_{ij} = 0$ for all i and j. Fig. 2 shows the traces of $\left\{ r_{ij}^{(t)} \right\}_{i,j=1,2}$ and $\left\{ \dot{r}_{ij}^{(t)} \right\}_{i,j=1,2}$ of the data sets (a)–(e) for each iteration with $\delta = 10^{-6}$, and illustrates that the sequence $\left\{ \dot{\theta}^{(t)} \right\}_{t\geqslant 0}$ converges to θ^{MLE} faster than $\left\{ \theta^{(t)} \right\}_{t\geqslant 0}$.

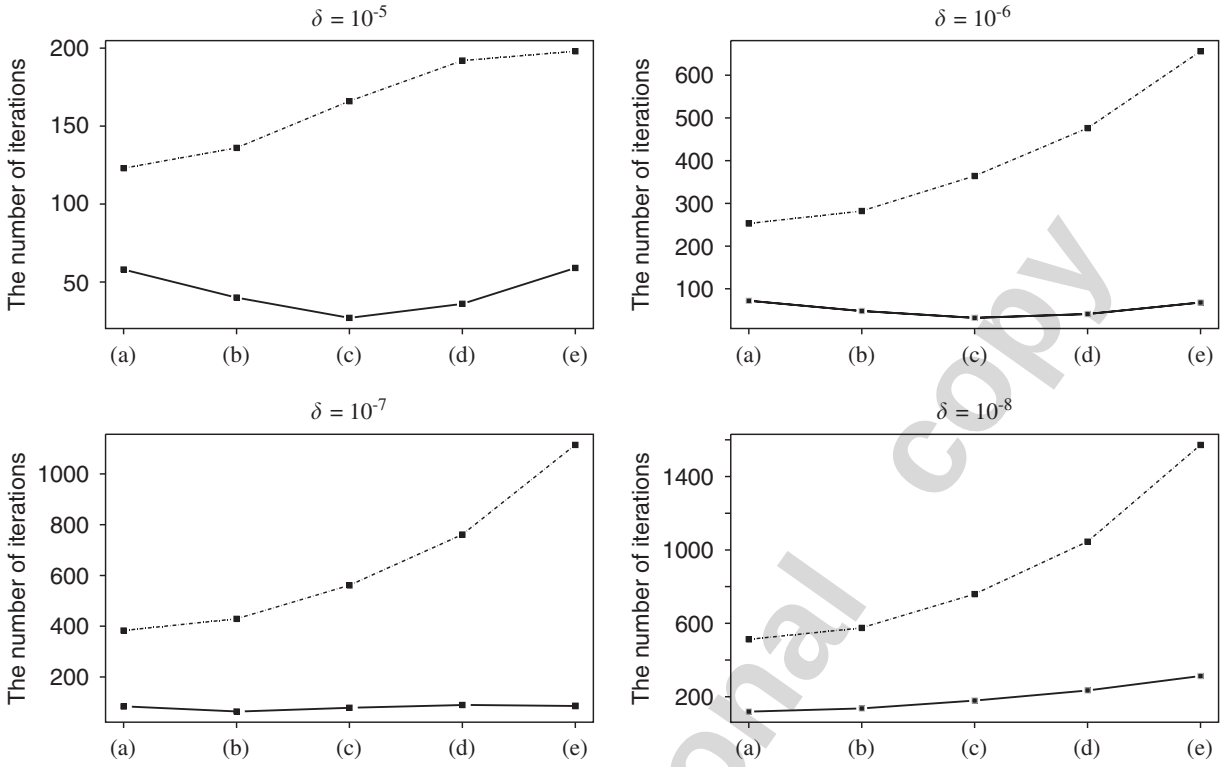

Fig. 1. Plots of the number of iterations for the EM (dashed line) and ε-accelerated EM (solid line) algorithms for each δ.

Table 3
Estimates of $\theta = \left\{p_{ij}\right\}_{i,j=1,2}$ by the EM and ε-accelerated EM algorithms for each δ

		$\delta = 10^{-5}$		$\delta = 10^{-6}$		$\delta = 10^{-7}, 10^{-8}$
		EM	ε-accelerated EM	EM	ε-accelerated EM	EM and ε-accelerated EM
(a)	p_{11}	0.3463	0.3457	0.3458	0.3458	0.3458
	p_{12}	0.2572	0.2577	0.2576	0.2577	0.2577
	p_{21}	0.2756	0.2761	0.2761	0.2761	0.2761
	p_{22}	0.1210	0.1205	0.1205	0.1204	0.1204
(b)	p_{11}	0.3471	0.3464	0.3465	0.3465	0.3465
	p_{12}	0.2564	0.2570	0.2569	0.2570	0.2570
	p_{21}	0.2763	0.2769	0.2768	0.2769	0.2769
	p_{22}	0.1203	0.1197	0.1197	0.1197	0.1197
(c)	p_{11}	0.3478	0.3469	0.3470	0.3469	0.3469
	p_{12}	0.2557	0.2565	0.2564	0.2565	0.2565
	p_{21}	0.2765	0.2774	0.2773	0.2774	0.2774
	p_{22}	0.1200	0.1192	0.1193	0.1192	0.1192
(d)	p_{11}	0.3483	0.3471	0.3472	0.3471	0.3471
	p_{12}	0.2551	0.2563	0.2562	0.2563	0.2564
	p_{21}	0.2763	0.2775	0.2774	0.2775	0.2776
	p_{22}	0.1202	0.1190	0.1191	0.1190	0.1190
(e)	p_{11}	0.3491	0.3472	0.3474	0.3472	0.3472
	p_{12}	0.2543	0.2563	0.2561	0.2563	0.2563
	p_{21}	0.2757	0.2776	0.2775	0.2776	0.2776
	p_{22}	0.1209	0.1189	0.1191	0.1190	0.1189

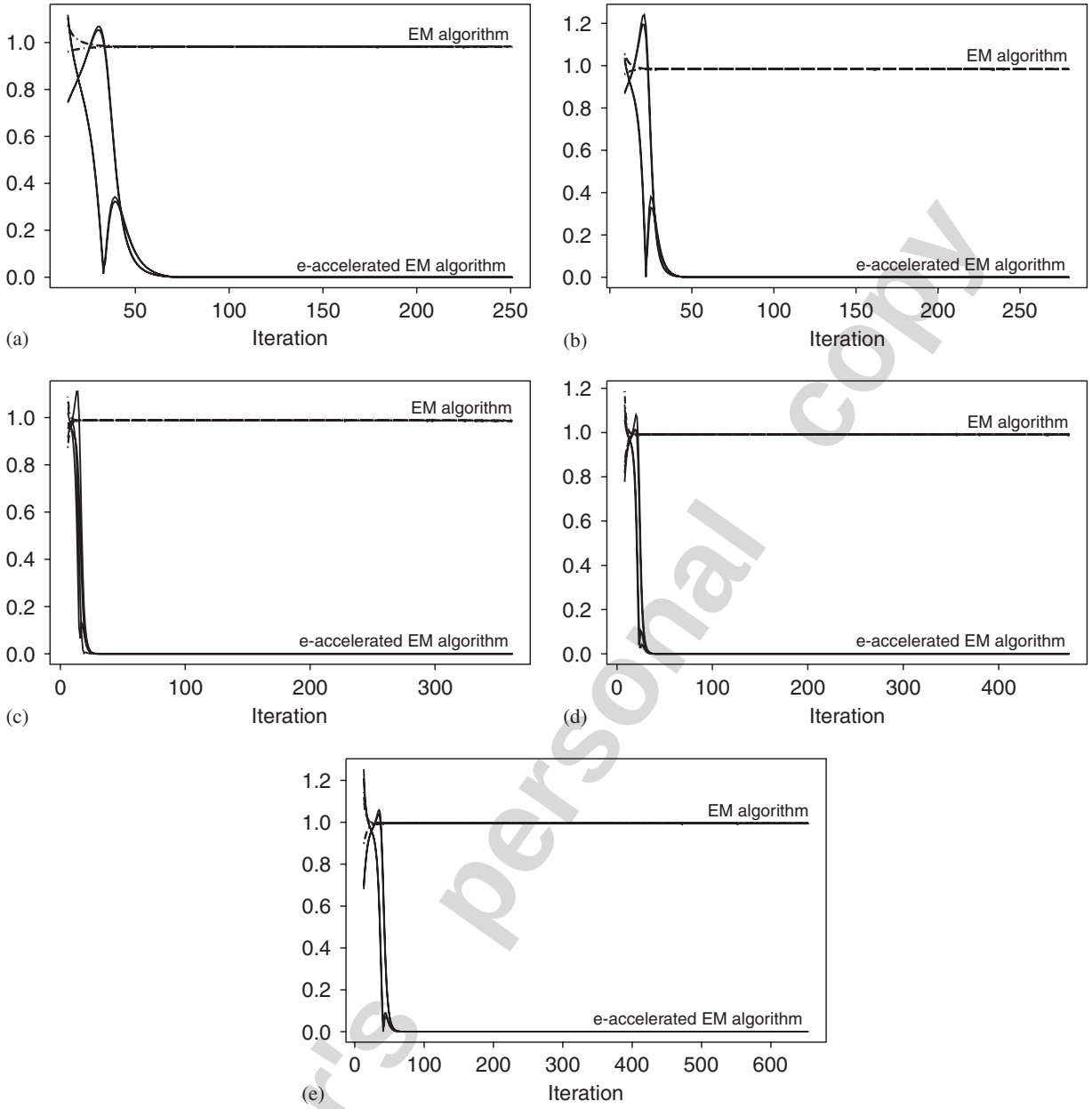

Fig. 2. Traces of $\left\{r_{ij}^{(t)}\right\}_{i,j=1,2}$ and $\left\{\dot{r}_{ij}^{(t)}\right\}_{i,j=1,2}$ for $\delta = 10^{-6}$ and the data sets (a)–(e).

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 observed data are shown in Table 4, where the observed data from No. 1 to 3 are complete, and the data from No. 4, 5 and No. 6, 7 are incomplete with missing X_2 and X_1, respectively.

The EM algorithm fills in missing data of the incompletely observed data in the E-step and estimates μ and Σ in the M-step. For the data set (a) and $\delta = 10^{-6}$, after 317 iterations of the EM algorithm, the MLEs of μ and Σ are obtained as follows:

$$\mu = (1.3005, 1.4163), \quad \Sigma = (0.2371, 4.9603, -1.0478).$$

The ε-accelerated EM algorithm finds the same values after 172 iterations using 174 EM iterations. Using both algorithms with the data set (b) and $\delta = 10^{-6}$, we obtain the same MLEs of μ and Σ:

$$\mu = (78.3977, 2247.1084), \quad \Sigma = (70.1051, 79869.7113, 2182.2234).$$

In this case the ε-accelerated EM algorithm takes 133 iterations, while the EM algorithm requires 312 iterations.

Table 4
Incomplete bivariate normal data

No.	X_1	X_2
(a)		
1	1.2	2.3
2	1.7	0.1
3	1.6	−0.7
4	0.2	—
5	1.5	—
6	—	−0.2
7	—	1.6
(b)		
1	68	2000
2	71	1850
3	72	2100
4	84	—
5	90	—
6	—	2150
7	—	2600

From these numerical experiments, we can also see that the sequence generated by the ε-accelerated EM algorithm converges to the MLEs faster than the EM sequence, as for the results of Example 1.

6. Concluding remarks

In this paper, we proposed the ε-accelerated EM algorithm that speeds up the convergence of the EM sequence incorporating the vector ε algorithm. The vector ε algorithm in itself is a fairly simple computational procedure and its computational cost is less expensive than that of Newton-type algorithms. The ε-accelerated EM algorithm does not require computation of the matrix inversion at each iteration and thus is numerically stable. Therefore, the ε-accelerated EM algorithm is an extension algorithm within the framework of the EM algorithm without affecting its simplicity, stability and flexibility.

Theoretically we showed that the ε-accelerated EM algorithm is guaranteed to converge to the stationary point of the EM sequence and that it speeds up the convergence of the scalar EM sequence. Numerical experiments demonstrate that the ε-accelerated EM algorithm produces a sequence that converges to MLEs much faster than the sequence generated by the EM algorithm. Hence, the ε-accelerated EM algorithm finds sufficiently accurate estimates using a smaller number of EM iterations.

In the future we intend to evaluate theoretically the speed of convergence of the ε-accelerated EM algorithm for vector cases.

Acknowledgement

The authors would like to thank the editor and two referees whose valuable comments and kindly suggestions that led to an improvement of this paper, and Zhi Geng and Mingfeng Wang for helpful discussion on the revision of this paper. This research is supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Young Scientists, no. 16700264, and the Wesco Scientific Promotion Foundation.

Appendix A

We here give the following equations used in the proof of Theorem 1.

Lemma 3. *For $t \to \infty$, we have*

$$\left\langle \theta^{(t+1)} - \theta^*, \theta^{(t)} - \theta^* \right\rangle = \{\lambda + \mathrm{o}(1)\} \left\| \theta^{(t)} - \theta^* \right\|^2, \tag{11}$$

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

$$\left\| \theta^{(t+1)} - \theta^{(t)} \right\|^2 = \left\{ (1-\lambda)^2 + \mathrm{o}(1) \right\} \left\| \theta^{(t)} - \theta^* \right\|^2. \tag{13}$$

Proof. We show in turn Eqs. (11)–(13).

Proof of (11): From Eq. (6), we have

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

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

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

so that

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

Thus, we can obtain

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

Proof of (12): For $t \to \infty$, we have

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

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

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

and then

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

Thus, we can obtain

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

Proof of (13): We have

$$\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2=\left\|\theta^{(t+1)}-\theta^*\right\|^2+\left\|\theta^{(t)}-\theta^*\right\|^2-2\left\langle\theta^{(t+1)}-\theta^*,\theta^{(t)}-\theta^*\right\rangle.$$

From Eqs. (11) and (12), we can obtain

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

Appendix B

Proof of Theorem 1. Set $\eta^{(t)}=\left[\theta^{(t)}-\theta^{(t+1)}\right]^{-1}+\left[\theta^{(t+2)}-\theta^{(t+1)}\right]^{-1}$. From Eq. (4), we have

$$\dot\theta^{(t)}-\theta^{(t+1)}=\left[\left[\theta^{(t)}-\theta^{(t+1)}\right]^{-1}+\left[\theta^{(t+2)}-\theta^{(t+1)}\right]^{-1}\right]^{-1}=\left[\eta^{(t)}\right]^{-1}\tag{14}$$

and thus we obtain

$$\left\langle\left(\dot\theta^{(t)}-\theta^{(t+1)}\right),\eta^{(t)}\right\rangle=1.\tag{15}$$

Here, we prove that $\eta^{(t)}$ diverges to infinity as t tends to infinity. From Eq. (6), we have

$$\begin{aligned}\theta^{(t+2)}-\theta^{(t+1)}&=\lambda\left(\theta^{(t+1)}-\theta^*\right)+\mathrm{O}\left(\left\|\theta^{(t+1)}-\theta^*\right\|^2\right)-\left\{\lambda\left(\theta^{(t)}-\theta^*\right)+\mathrm{O}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right)\right\}\\&=\lambda\left(\theta^{(t+1)}-\theta^{(t)}\right)+\mathrm{O}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right),\quad t\to\infty.\end{aligned}\tag{16}$$

For $t\to\infty$, we have

$$\begin{aligned}\eta^{(t)}&=\frac{\theta^{(t)}-\theta^{(t+1)}}{\left\|\theta^{(t)}-\theta^{(t+1)}\right\|^2}+\frac{\theta^{(t+2)}-\theta^{(t+1)}}{\left\|\theta^{(t+2)}-\theta^{(t+1)}\right\|^2}\\&=\frac{\theta^{(t)}-\theta^{(t+1)}}{\left\|\theta^{(t)}-\theta^{(t+1)}\right\|^2}+\frac{\lambda\left(\theta^{(t+1)}-\theta^{(t)}\right)+\mathrm{O}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right)}{\left\|\theta^{(t+2)}-\theta^{(t+1)}\right\|^2}\\&=\left\{\frac{-1}{\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2}+\frac{\lambda}{\left\|\theta^{(t+2)}-\theta^{(t+1)}\right\|^2}\right\}\left(\theta^{(t+1)}-\theta^{(t)}\right)+\frac{\mathrm{O}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right)}{\left\|\theta^{(t+2)}-\theta^{(t+1)}\right\|^2}.\end{aligned}$$

From Lemma 3 and Eq. (16), we obtain

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

because $\lim_{t\to\infty}\theta^{(t)} = \lim_{t\to\infty}\theta^{(t+1)} = \theta^*$. Then, from Eq. (13), we have

$$\begin{aligned}\frac{-1}{\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2}+\frac{\lambda}{\left\|\theta^{(t+2)}-\theta^{(t+1)}\right\|^2} &= \frac{-1}{\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2}+\frac{\lambda}{\lambda^2\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2+\mathrm{o}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right)}\\ &= \frac{\lambda(1-\lambda)\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2+\mathrm{o}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right)}{\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2\left\{\lambda^2\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2+\mathrm{o}\left(\left\|\theta^{(t)}-\theta^*\right\|^2\right)\right\}}\\ &= \frac{\lambda(1-\lambda)^3+\mathrm{o}(1)}{\lambda^2(1-\lambda)^2+\mathrm{o}(1)}\frac{1}{\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2}, \quad t\to\infty,\end{aligned}$$

and, from Eqs. (12) and (13), we also have

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

Thus, for $t\to\infty$, we can obtain

$$\eta^{(t)} = \frac{\lambda(1-\lambda)^3+\mathrm{o}(1)}{\lambda^2(1-\lambda)^2+\mathrm{o}(1)}\frac{\theta^{(t+1)}-\theta^{(t)}}{\left\|\theta^{(t+1)}-\theta^{(t)}\right\|^2}+\frac{\mathrm{O}(1)}{\lambda^2(1-\lambda)^2+\mathrm{o}(1)}.$$

Since $\lim_{t\to\infty}\theta^{(t)} = \lim_{t\to\infty}\theta^{(t+1)} = \theta^*$, $\eta^{(t)}$ diverges to infinity as $t\to\infty$. For any t, Eq. (15) holds, so that $\lim_{t\to\infty}\left(\dot\theta^{(t)}-\theta^{(t+1)}\right) = \vec{0}$, where $\vec{0}$ is the zero vector.

Next we consider the case that some element of $\theta^{(t)}$ has converged, that is, $\theta_i^{(t)}-\theta_i^* = 0$ for some $i\in\{1,2,\ldots,d\}$ and all t that are larger than some constant M. Let $\eta_i^{(t)}$ denote the ith element of $\eta^{(t)}$. Then, for all $t > M+2$, we obtain $\dot\theta_i^{(t)}-\theta_i^{(t+1)} = 0$ from Eq. (14). Since $\lim_{t\to\infty}\theta_i^{(t+1)} = \theta_i^*$, we have

$$\lim_{t\to\infty}\left(\dot\theta_i^{(t)}-\theta_i^*\right) = 0.$$

From Eq. (15), we also have, for any t,

$$\left\langle\left(\dot\theta^{(t)}-\theta^{(t+1)}\right),\eta^{(t)}\right\rangle = \sum_{j\neq i}\left(\dot\theta_j^{(t)}-\theta_j^{(t+1)}\right)\eta_j^{(t)} = 1.$$

Since each $\eta_j^{(t)}$ $(j\neq i)$ diverges to infinity as $t\to\infty$, we can obtain

$$\lim_{t\to\infty}\left(\dot\theta_j^{(t)}-\theta_j^{(t+1)}\right) = \lim_{t\to\infty}\left(\dot\theta_j^{(t)}-\theta_j^*\right) = 0. \qquad \square$$

References

Aitken, A.C., 1926. On Bernoulli's numerical solution of algebraic equations. Proc. Roy. Soc. Edinburg 46, 289–305.

Brezinski, C., Zaglia, M.R., 1991. Extrapolation Methods: Theory and Practice. Elsevier Science Ltd. North-Holland, Amsterdam.

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

Geng, Z., Wan, K., Tao, F., 2000. Mixed graphical models with missing data and the partial imputation EM algorithm. Scand. J. Statist. 27, 433–444.

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

Lange, K., 1995. Acceleration of the EM algorithm by using Quasi-Newton methods. J. Roy. Statist. Soc. Ser. B 59, 569–587.

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

Louis, T.A., 1982. Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44, 226–233.
McLachlan, G.J., Krishnan, T., 1997. The EM Algorithm and Extensions. Wiley, New York.
Meilijson, I., 1989. A fast improvement to the EM algorithm on its own terms. J. Roy. Statist. Soc. Ser. B 51, 127–138.
Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267–278.
Meng, X.L., Rubin, D.B., 1994. On the global and componentwise rates of convergence of the EM algorithm. Linear Algebra Appl. 199, 413–425.
Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman & Hall, London.
Traub, J.F., 1964. Iterative Methods for the Solution of Equations. Prentice-Hall Inc., Englewood Cliffs, NJ.
Wynn, P., 1961. The epsilon algorithm and operational formulas of numerical analysis. Math. Comp. 15, 151–158.
Wynn, P., 1962. Acceleration techniques for iterated vector and matrix problems. Math. Comp. 16, 301–322.