# Acceleration of the EM and ECM algorithms for log-linear models with missing data

Masahiro Kuroda[1] and Michio Sakakihara[2]

[1] Okayama University of Science
Japan
kuroda@soci.ous.ac.jp
[2] Okayama University of Science
Japan
sakaki@mis.ous.ac.jp

**Summary.** The EM algorithm has been a general and popular algorithm for finding maximum likelihood estimates (MLEs) from incomplete data since Dempster, Laird and Rubin [DLR77]. However, it is often criticized that the convergence of the EM algorithm is slow when the proportion of missing data is high. Kuroda and Sakakihara [KS05] proposed the $\varepsilon$-accelerated EM algorithm that is the fairly simple computational procedure and speeds up the convergence of EM sequences using the vector $\varepsilon$ algorithm of Wynn [Wy62]. In this paper, we use the $\varepsilon$-accelerator for the EM and ECM algorithms for log-linear models with missing data. Moreover we apply the Aitken $\delta^2$ method of Aitken [A26] to these algorithms. The Aitken $\delta^2$ method also speeds up the convergence of EM sequences. Then we provide the convergence properties of both accelerators.

**Key words:** vector $\varepsilon$ algorithm, Aitken $\delta^2$ method, acceleration of convergence, EM algorithm, ECM algorithm, log-linear models, missing data

## 1 Introduction

The EM algorithm has been a general and popular algorithm for finding maximum likelihood estimates (MLEs) from incomplete data since Dempster, Laird and Rubin [DLR77]. However, it is often criticized that the convergence of the EM algorithm is slow when the proportion of missing data is high. In order to improve the speed of convergence of the EM algorithm, various algorithms incorporating optimization algorithms with faster convergence speed have been proposed, see Krishnan and McLachlan [MK97]. Applying the accelerator based on the Newton-type algorithms, it requires the computation of a matrix inversion at each iteration. Then its computation is likely to become rapidly complicated as the number of parameters increases, and is also expected numerical instabilities. Therefore the Newton-type accelerators for the EM algorithm are lost the attractive features of the EM algorithm, such as its stability, flexibility and simplicity.

Kuroda and Sakakihara [KS05] proposed the $\varepsilon$-accelerated EM algorithm that is the fairly simple computational procedure and speeds up the convergence of EM sequences using the vector $\varepsilon$ algorithm of Wynn [Wy62], and demonstrated that the algorithm produces a sequence to converge to the MLEs much faster than the EM sequence in the numerical experiments.

In this paper, we apply the $\varepsilon$-accelerated EM algorithm to the maximum likelihood estimation for log-linear models with missing data. Moreover we propose the acceleration of the EM algorithm using the Aitken $\delta^2$ method of Aitken [A26] to accelerate the convergence of scalar sequences of iterates. For some log-linear models, there are no closed-form to compute the MLEs. The ECM algorithm of Meng and Rubin [MR93] is also used to the models. Then the $\varepsilon$- and Aiken $\delta^2$-accelerators are applicable to the ECM algorithms.

In Section 2 we introduce log-linear models for contingency tables, and, in Section 3, show the EM and ECM algorithms for the log-linear models with missing data. In Section 4, we provide the the $\varepsilon$- and Aitken $\delta^2$-accelerations for the EM and ECM algorithms, and give the fundamental theories for the convergence of these acceleration algorithms.

## 2 Log-linear models with missing data

Let $X_V = (X_1, \ldots, X_k)$ be a $k$-dimensional discrete random vector and $V = \{1, \ldots, k\}$ be the index set of $X_V$. We also denote a finite set of values of $X_i$ as $\Omega_i$, $i \in V$, and the space of possible values of $X_V$ as the Cartesian product $\Omega_V = \prod_{i \in V} \Omega_i$. Associated with each $i \in V$, we shall have a random variable $X_i$ taking values in a sample space $\Omega_i$. For a subset $A \subseteq V$, we write $X_A$ for $\{X_i | i \in A\}$ and $\Omega_A = \prod_{i \in A} \Omega_i$.

Let $p_V(x_V)$ denote the cell probability of $X_V = x_V \in \Omega_V$ and let $\theta = \{p_V(x_V) | x_V \in \Omega_V\}$ be the set of cell probabilities. The marginal probability of $X_A = x_A \in \Omega_A$ for $A \subset V$ can be also calculated by

$$p_A(x_A) = \sum_{x_{V \setminus A}} p_V(x_V),$$

where the symbol "\" denotes the operator of a difference set.

In this paper, we consider the class of hierarchical log-linear models for contingency tables in the case of multinomial sampling. An hierarchical log-linear model is specified by a generating class $E = \{e_1, \ldots, e_M\}$ which is a class of variable sets in minimal interaction terms, and $e \in E$ is called a generator. For example, consider the log-linear model with the generating class $E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$. Then the log-linear model is represented by

$$\begin{aligned} \log p_V(x_V) &= u + u_1(x_{\{1\}}) + u_2(x_{\{2\}}) + u_3(x_{\{3\}}) \\ &\quad + u_{12}(x_{\{1,2\}}) + u_{12}(x_{\{1,3\}}) + u_{23}(x_{\{2,3\}}). \end{aligned} \tag{1}$$

In the cases with missing data, the completely classified data are represented in the full contingency tables and the partially classified data are represented in supplemental tables. These tables are identified with an observing pattern $T =$

$\{V, t_1, \ldots, t_S\}$ where each configuration $t$ is a set of observed variables for a supplemental table. The cell counts of the contingency table classified by observing variables $X_t$ are also denoted as $\{n_t(x_t)|x_t \in \Omega_t\}$. Assume that the missing data are missing at random in the sense that Rubin [R76].

## 3 The EM and ECM algorithms for log-linear models with missing data

We show the EM algorithm to find the MLE of the parameter vector $\theta$ for the log-linear model with a generating class $E$ and an observing pattern $T = \{V, t_1, \ldots, t_S\}$.

For the case that the MLE of $\theta$ is given by closed-forms, the EM algorithm repeats the following two steps till the convergence to the desired accuracy is obtained:

- E-step: Calculate the expected marginal cell counts $\{\tilde{n}_e(x_e)\}$ for each generator $e \in E$:

$$\tilde{n}_e(x_e)^{(r)} = \sum_{x_{V \setminus e}} \left[ n_V(x_V) + \sum_{i=1}^{S} \frac{p_V(x_V)^{(r)}}{p_{t_i}(x_{t_i})^{(r)}} n_{t_i}(x_{t_i}) \right].$$

- M-step: Calculate marginal probabilities $\{p_e(x_e)^{(r+1)}|x_e \in \Omega_e\}$ for each generator $e \in E$:

$$p_e(x_e)^{(r+1)} = \frac{\tilde{n}_e(x_e)^{(r)}}{N},$$

where $N = \sum_{t \in T} \sum_{x_t \in \Omega_t} n_t(x_t)$. And Obtain $\theta^{(r+1)}$ from

$$p_V(x_V)^{(r+1)} = \frac{\prod_{e \in E} p_e(x_e)^{(r+1)}}{\prod_{c \in C} (p_c(x_c)^{(r+1)})^{\nu(c)}},$$

where $C = (c_1, \ldots, c_{M-1})$ is the set of separators of $E$ and $\nu(c)$ is an index that counts the number of times of each separator $c$ in $C$.

It is known that certain log-linear models do not have closed-form MLEs, such as the no three-interaction model for a three-way contingency table given by equation (1). Then the ECM algorithm of Meng and Rubin [MR93] is applicable to the models. The ECM algorithm finds the MLEs by replacing a complicated M-step of the EM algorithm with several computationally simpler CM-step:

- CM-step: Obtain $\theta^{(r+1)}$ by performing the Iterative Proportional Fitting procedure of Bishop, Fienberg and Holland [BFH77] as follows:

$$p_V(x_V)^{(r+i/M)} = \frac{p_V(x_V)^{(r+(i-1)/M)}}{p_{e_i}(x_{e_i})^{(r+(i-1)/M)}} \frac{\tilde{n}_{e_i}(x_{e_i})^{(r)}}{N},$$

for $i = 1, 2, \ldots, M$.

## 4 Acceleration algorithms

In order to accelerate the convergence of the EM and ECM algorithms, we incorporate the $\varepsilon$- and Aitken $\delta^2$-accelerators into them. The first accelerator is the vector $\varepsilon$ algorithm of Wynn [Wy62] and accelerates the convergence of *vector sequences*. The second accelerator is the Aitken $\delta^2$ method of Aitken [A26] and accelerates the convergence of *scalar sequences*.

### 4.1 The $\varepsilon$ acceleration

Let $\theta$ be a $d$-dimensional vector. Assume that a sequence $\{\theta^{(r)}\}_{r\geq 0}$ converges to a vector $\theta^*$ as $r \to \infty$. Define the inverse $[x]^{-1}$ of a vector $x$ by

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

in which $< x, x >$ is the scalar product of $x$ by itself.

In general, the vector $\varepsilon$ algorithm for a sequence $\{\theta^{(r)}\}_{r\geq 0}$ starts with

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

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

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

For the case of $k + 1 = 2l + 2$, we have the iteration form such as

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

from equation (2), see Brezinski and Zaglia [BZ91]. For practical implementation, we apply the case of $l = 0$ to equation (**??**). Then, from initial conditions $\varepsilon^{(r,0)} = \theta^{(r)}$ and $\varepsilon^{(r,-2)} = \infty$ of Brezinski and Zaglia [BZ91], the iteration becomes as follows:

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

To accelerate the convergence of the EM and ECM algorithms, we apply the $\varepsilon$-acceleration process to the sequence $\{\theta^{(t)}\}_{t\geq 0}$ generated by the M- or CM-steps:

- $\underline{\varepsilon\text{-acceleration:}}$ Generate a vector $\dot{\theta}^{(r)}$ by

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

  from equation (1) and check the convergence to a desired accuracy.

Then the acceleration algorithm via the vector $\varepsilon$ algorithm does not improve the E- and M-steps or the CM-step but accelerates the convergence of the sequence $\{\theta^{(r)}\}_{r\geq 0}$ using the $\varepsilon$-acceleration process.

### 4.2 Convergence of the $\varepsilon$-acceleration process

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

$$\theta^{(r)} = M(\theta^{(r-1)}).$$

Expanding $M(\theta^{(r)})$ in a first term of a Taylor series of about $\theta^{(r-1)}$, the following approximation holds:

$$\theta^{(r+1)} - \theta^{(r)} = M(\theta^{(r)}) - M(\theta^{(r-1)}) \doteq J^{(r-1)}(\theta^{(r)} - \theta^{(r-1)}),$$

where $J^{(r-1)}$ is the Jacobian matrix for the mapping $M(\theta)$ evaluated at $\theta^{(r-1)}$.

Suppose $\theta^{(r)}$ converges to a stationary point $\theta^*$. For $r$ large enough, we have $J^{(r)} = J^*$ and thus

$$\theta^{(r+1)} - \theta^{(r)} \doteq J^*(\theta^{(r)} - \theta^{(r-1)}), \tag{4}$$

where $J^*$ is $J$ evaluated at a stationary point $\theta^*$. Then $J^*$ determines the rate of convergence of the EM algorithm near $\theta^*$. In fact, equation (4) becomes

$$\theta^{(r+1)} - \theta^{(r)} \doteq \lambda(\theta^{(r)} - \theta^{(r-1)}), \tag{5}$$

because the largest eigenvalue $\lambda$ of $J^*$ dominates the convergence, see Louis [L82].

Kuroda and Sakakihara [KS05] provided the convergence of the $\varepsilon$-acceleration process.

**Theorem 1.** *The sequence $\{\dot{\theta}^{(r)}\}_{r\geq 0}$ generated by the $\varepsilon$-acceleration process for the EM sequence converges to the stationary point $\theta^*$ of the EM sequence.*

Like the EM algorithm, equation (5) also holds for each iteration of the ECM algorithm when $r$ tends to be large. Thus Theorem 1 is valid for ECM sequences.

**Corollary 1.** *The sequence $\{\dot{\theta}^{(r)}\}_{r\geq 0}$ generated by the $\varepsilon$-acceleration process for the ECM sequence converges to the stationary point $\theta^*$ of the ECM sequence.*

For the convergence and acceleration properties of the vector $\varepsilon$ algorithm, Brezinski and Zaglia [BZ91] has completely described only for special classes of vector sequences. Thus the mathematical consideration of the $\varepsilon$-acceleration process is less trivial.

### 4.3 The Aitken $\delta^2$-acceleration

Let a scalar sequence $\{\phi^{(r)}\}_{r\geq 0}$ converges to a limit $\phi^*$. For a scalar sequence $\{\phi^{(r)}\}_{r\geq 0}$, the Aitken $\delta^2$ algorithm generates a sequence $\{\dot{\phi}^{(r)}\}_{r\geq 0}$ by

$$\dot{\phi}^{(r)} = \phi^{(r)} - \frac{(\phi^{(r+1)} - \phi^{(r)})^2}{\phi^{(r+2)} - 2\phi^{(r+1)} + \phi^{(r)}}. \tag{6}$$

Note that the $\varepsilon$ algorithm for scalar sequences is identical to the Aitken $\delta^2$ algorithm.

For the sake of simplicity, we re-denote $\theta = \{p_V(x_V)|x_V \in \Omega_V\}$ as $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$ where $\theta_d = 1 - \sum_{j=1}^{d-1} \theta_j$. We can find an alternative parameterization $\varphi = \varphi(\theta)$ such as

$$\varphi = (\phi_1, \phi_2, \ldots, \phi_d) = \left(\theta_1, \frac{\theta_2}{1-\theta_1}, \ldots, \frac{\theta_{d-1}}{1-(\sum_{j<d-1}\theta_j)}, \frac{\theta_d}{1-(\sum_{j<d-1}\theta_j)}\right). \tag{7}$$

Then the MLE of $\varphi(\theta)$ is $\varphi(\hat{\theta})$ that is the function evaluated at the MLE $\hat{\theta}$ of $\theta$, because $\phi(\theta)$ is a one-one function of $\theta$. For the parameterization of $\varphi = \varphi(\theta)$ in equation (7), the property of the invariance of MLEs holds.

In order to accelerate the convergence of the EM or ECM sequence $\{\theta^{(r)}\}_{r\geq 0}$, the Aitken $\delta^2$-acceleration process generates a sequence $\{\dot{\varphi}^{(r)}\}_{r\geq 0}$:

- Aitken $\delta^2$-acceleration: Generate a vector $\dot{\varphi}^{(r)} = (\dot{\phi}_i^{(r)})_{i=1,\ldots,d}$ from

$$\dot{\phi}_i^{(r)} = \phi_i^{(r)} - \frac{(\phi_i^{(r+1)} - \phi_i^{(r)})^2}{\phi_i^{(r+2)} - 2\phi_i^{(r+1)} + \phi_i^{(r)}}, \qquad i = 1, 2, \ldots, d,$$

  and check the convergence to a desired accuracy.

After the convergence of the Aitken $\delta^2$ acceleration, we compute $\dot{\theta}$ from the limit $\dot{\varphi}^*$ of $\{\dot{\varphi}^{(r)}\}_{r\geq 0}$ using

$$\begin{aligned}\dot{\theta} &= (\dot{\theta}_1, \dot{\theta}_2, \ldots, \dot{\theta}_{d-1}, \dot{\theta}_d) \\ &= \left(\dot{\phi}_1^*, \dot{\phi}_2^*(1-\dot{\phi}_1^*), \ldots, \dot{\phi}_{d-1}^* \prod_{j=1}^{d-2}(1-\dot{\phi}_j^*), \dot{\phi}_d^* \prod_{j=1}^{d-2}(1-\dot{\phi}_j^*)\right).\end{aligned}$$

### 4.4 Convergence of the Aitken $\delta^2$-acceleration process

Next we give the convergence properties of the Aitken $\delta^2$ acceleration for the EM and ECM algorithms. For a scalar sequence $\{\dot{\phi}^{(r)}\}_{r\geq 0}$ generated by equation (6), Traub [T64] provided the following lemma:

**Lemma 1.** *If $\{\phi^{(r)}\}_{r\geq 0}$ converges to a stationary point $\phi^*$ as $r \to \infty$, then $\{\dot{\phi}^{(r)}\}_{r\geq 0}$ generated by equation (6) converges to the same stationary point $\phi^*$.*

Then we have the following lemma:

**Lemma 2.** *If $\{\varphi^{(r)}\}_{r\geq 0}$ is a convergent vector sequence to the vector $\varphi^*$, then $||\varphi^{(r)} - \varphi^*||_\infty$ converges to zero as $r \to \infty$, moreover $|\phi_i^{(r)} - \phi_i^*|$ converges to zero as $r \to \infty$, where $\|\varphi\|_\infty = \max_i\{|\phi_i|\}$.*

From Lemmas 1 and 2, we obtain the result:

**Theorem 2.** *If $\varphi^{(r)} = (\phi_i^{(r)})_{i=1,\ldots,d}$ is a convergent vector sequence to the vector $\varphi^* = (\phi_i^*)_{i=1,\ldots,d}$ and the vector sequence $\dot{\varphi}^{(r)} = (\dot{\phi}_i^{(r)})_{i=1,\ldots,d}$ is generated by equation (6), then $\dot{\varphi}^{(r)} \to \varphi^*$ as $r \to \infty$.*

Furthermore, from the invariance of MLEs of $\varphi = \varphi(\theta)$ in (7) and Theorem 2, we can give the following result:

**Theorem 3.** *If the EM sequence $\{\theta^{(r)}\}_{r\geq 0}$ converges to the MLE $\hat{\theta}$, then $\dot{\theta}$ obtained by equation (**??**) is $\hat{\theta}$.*

To compare the speed of convergence of the Aitken $\delta^2$ acceleration for the EM algorithm with that of the EM algorithm, we provide the following notion of Brezinski and Zaglia [BZ91].

**Definition 1.** *Let* $\{\hat{\phi}^{(r)}\}_{r\geq 0}$ *be a scalar sequence obtained by applying an extrapolation method to* $\{\phi^{(r)}\}_{r\geq 0}$*. Assume that* $\lim_{t\to\infty}\phi^{(r)} = \lim_{t\to\infty}\hat{\phi}^{(r)} = \phi^*$*. If*

$$\lim_{t\to\infty}\frac{|\hat{\phi}^{(r)} - \phi^*|}{|\phi^{(r)} - \phi^*|} = 0,$$

*then we say that the sequence* $\{\hat{\phi}^{(r)}\}_{r\geq 0}$ *converges to* $\phi^*$ *faster than* $\{\phi^{(r)}\}_{r\geq 0}$ *or the extrapolation method accelerates the convergence of* $\{\phi^{(r)}\}_{r\geq 0}$*.*

Traub [T64] proved that Aitken $\delta^2$ method accelerates the convergence of $\{\phi^{(r)}\}_{r\geq 0}$ in the sense that

$$\lim_{r\to\infty}\frac{|\dot{\phi}^{(t)} - \phi^*|}{|\phi^{(t+2)} - \phi^*|} = 0. \tag{8}$$

The facts of Theorem 3 and equation (8) give us the validity of the Aitken $\delta^2$ acceleration for the EM algorithm.

## Acknowledgement

The authors would like to thank a referee for valuable comments. 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.

## References

[A26] Aitken, A.C.: On Bernulli's numerical solution of algebraic equations. Prc. R. Soc. Edinb., 46, 623-634 (1926)

[BFH77] Bishop, Y.M.M., Fienberg, S.E. and Holland, P.W.: Discrete multivariate analysis: theory and practice. The M.I.T. Press, Cambridge, Mass.-London (1977)

[BZ91] Brezinski, C. and Zaglia, M.R.: Extrapolation methods: theory and practice. Elsevier Science Ltd. North-Holland, Amsterdam (1991)

[DLR77] Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B 39, 1-22 (1977)

[KS05] Kuroda, M. and Sakakihara, M.: Accelerating the convergence of the EM algorithm using the vector $\varepsilon$ algorithm. *in review process* (2005)

[L82] Louis, T.A.: Finding the observed information matrix when using the EM algorithm. J. Roy. Statist. Soc. Ser. B 44, 226-233 (1982)

[MK97] McLachlan, G.J. and Krishnan, T.: The EM algorithm and extensions. Wiley, New York (1997)

[MR93] Meng, X.L. and Rubin, D.B.: Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267-278 (1993)

[R76] Rubin, D.B.: Inference and missing data. Biometrika 63, 581-592 (1976)

[T64] Traub, J.F.: Iterative Methods for the Solution of Equations. Prentice-Hall, Inc., Englewood Cliffs, N.J. (1964)

[Wy62] Wynn, P.: Acceleration techniques for iterated vector and matrix problems. Math. Comp. 16, 301-322 (1962)