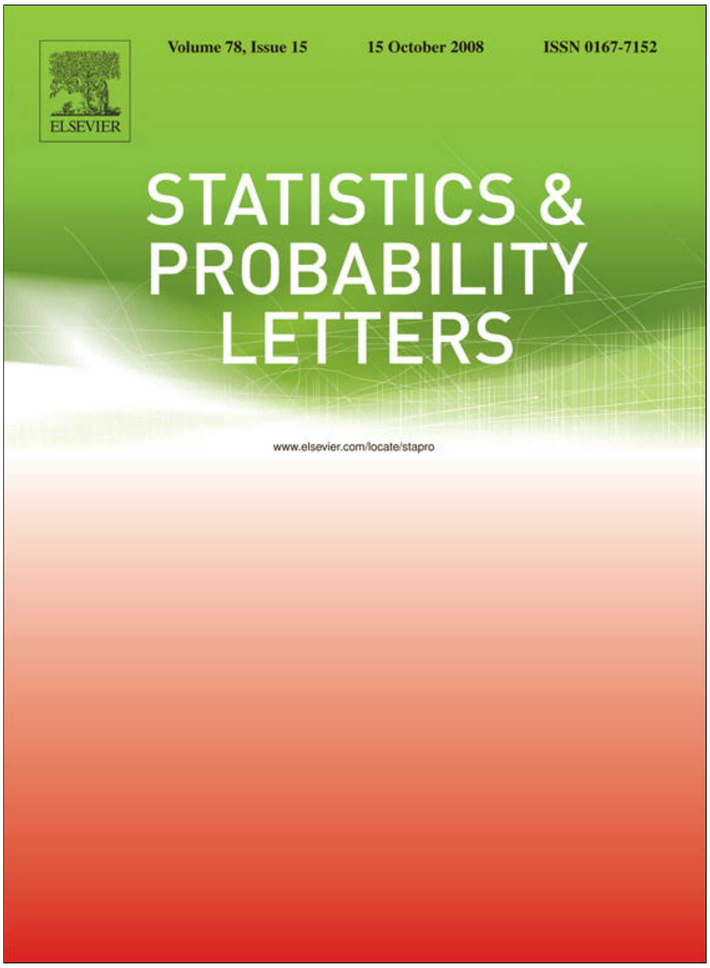

Volume 78, Issue 15 15 October 2008 ISSN 0167-7152

ELSEVIER

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Available online at www.sciencedirect.com

ScienceDirect

Statistics and Probability Letters 78 (2008) 2332–2338

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Acceleration of the EM and ECM algorithms using the Aitken δ^2 method for log-linear models with partially classified data

Masahiro Kuroda[a,*], Michio Sakakihara[b], Zhi Geng[c]

[a] *Department of Socio-Information, Okayama University of Science, 1-1 Ridaicho, Okayama 700-0005, Japan*
[b] *Department of Information Science, Okayama University of Science, 1-1 Ridaicho, Okayama 700-0005, Japan*
[c] *School of Mathematical Sciences, Peking University, Beijing 100871, China*

Received 14 August 2006; received in revised form 3 January 2008; accepted 21 January 2008
Available online 17 February 2008

Abstract

In this paper, we discuss the MLEs for log-linear models with partially classified data. We propose to apply the Aitken δ^2 method of Aitken [Aitken, A.C., 1926. On Bernoulli's numerical solution of algebraic equations. Proc. R. Soc. Edinburgh 46, 289–305] to the EM and ECM algorithms to accelerate their convergence. The Aitken δ^2 accelerated algorithm shares desirable properties of the EM algorithm, such as numerical stability, computational simplicity and flexibility in interpreting the incompleteness of data. We show the convergence of the Aitken δ^2 accelerated algorithm and compare its speed of convergence with that of the EM algorithm, and we also illustrate their performance by means of a simulation.

1. Introduction

The expectation and maximization (EM) algorithm proposed by Dempster et al. (1977) is often used to find maximization likelihood estimates (MLEs) for log-linear models with partially classified data because of its stability, flexibility and simplicity, see Fuchs (1982) and Chen et al. (1984). However the EM algorithm converges slowly when there is a relatively large proportion of partially classified data. Many algorithms have been proposed to speed up the convergence of the EM algorithm, see McLachlan and Krishnan (1997). Newton-type accelerators require the computation of a matrix inversion at each iteration of the EM algorithm. The computation becomes increasingly more complicated as the number of parameters increases, and it also becomes numerically unstable.

In this paper, we apply the Aitken δ^2 method of Aitken (1926) to the EM and ECM algorithms to speed up their convergence. The Aitken δ^2 acceleration does not affect the simplicity, stability and flexibility of the EM algorithm. The Aitken δ^2 accelerated method is then applied to log-linear models with partially classified data.

In Section 2, we introduce log-linear models and the EM and ECM algorithms for partially classified data. In Section 3, we provide the Aitken δ^2 acceleration for the EM and ECM algorithms and present theoretical results concerning their convergence. Numerical experiments in Section 4 illustrate the performance of Aitken δ^2 acceleration for the EM and ECM algorithms.

[*] Corresponding author.
E-mail address: kuroda@soci.ous.ac.jp (M. Kuroda).

0167-7152/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.01.102

2. Log-linear models and the EM algorithm

Let $X_V = (X_1, \dots, X_k)$ be a k-dimensional discrete random vector indexed by $V = \{1, \dots, k\}$. We also denote a finite set of values of X_i by Ω_i for $i \in V$, and the space of possible values of X_V as the Cartesian product $\Omega_V = \prod_{i \in V} \Omega_i$. 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 $p(x_V)$ denote the cell probability that $X_V = x_V$ and let $\theta = \{p(x_V) | x_V \in \Omega_V\}$ be the set of cell probabilities. The marginal probability that $X_A = x_A$ for $A \subset V$ can be calculated by

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

where the symbol "\" denotes the set difference.

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

$$\log p(x_V) = u + u_1(x_{\{1\}}) + u_2(x_{\{2\}}) + u_3(x_{\{3\}}) + u_{12}(x_{\{1,2\}}) + u_{12}(x_{\{1,3\}}) + u_{23}(x_{\{2,3\}}). \tag{1}$$

For the situation of 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 indexed with an observing pattern $T = \{V, t_1, \dots, t_S\}$, where each configuration t is a set of observed variables for a supplemental table. The cell counts of the contingency table classified by the observed variables X_t are denoted by $n_t = \{n_t(x_t) | x_t \in \Omega_t\}$. Assume that the missing data are missing at random in the sense of Rubin (1976).

Below we show the EM algorithm for finding the MLE of the parameter vector θ for a log-linear model with a generating class E and an observing pattern $T = \{V, t_1, \dots, t_S\}$. For log-linear models with closed-form MLEs, the EM algorithm repeats the following two steps until convergence to a desired accuracy is obtained:

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

$$\tilde{n}_e(x_e)^{(r)} = \sum_{x_{V \setminus e}} \left[n_V(x_V) + \sum_{i=1}^{S} \frac{p(x_V)^{(r)}}{p(x_{t_i})^{(r)}} n_{t_i}(x_{t_i}) \right]. \tag{2}$$

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

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

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

$$p(x_V)^{(r+1)} = \frac{\prod\limits_{e \in E} p(x_e)^{(r+1)}}{\prod\limits_{c \in C} (p(x_c)^{(r+1)})^{\nu(c)}}, \tag{3}$$

where $C = (c_1, \dots, c_{M-1})$ is a set of separators of E which may have the same elements, and $\nu(c)$ counts the number of separators equal to c. C can be obtained by

$$c_i = e_{\sigma(j)} \cap (e_{\sigma(1)} \cup \dots \cup e_{\sigma(j-1)}), \quad \text{for } 1 \le i < j \le M,$$

where the generating class $E = \{e_1, \dots, e_M\}$ can be ordered as $E = (e_{\sigma(1)}, e_{\sigma(2)}, \dots, e_{\sigma(M)})$ in such a way that it has the running intersection property:

$$e_{\sigma(j)} \cap (e_{\sigma(1)} \cup \dots \cup e_{\sigma(j-1)}) \subset e_{\sigma(i)},$$

for some $i < j$, see Lauritzen (1995).

When log-linear models do not have closed-form MLEs, such as in the case of the model (1), the ECM algorithm of Meng and Rubin (1993) may be applied. In the ECM algorithm, the M-step of the EM algorithm is replaced by several computationally simpler CM-steps:

CM-step: Find $\theta^{(r+1)}$ with the iterative proportional fitting procedure of Bishop et al. (1975) as follows:

$$p(x_V)^{(r+i/M)} = \frac{p(x_V)^{(r+(i-1)/M)}}{p(x_{e_i})^{(r+(i-1)/M)}} \frac{\tilde{n}_{e_i}(x_{e_i})^{(r)}}{N}, \tag{4}$$

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

At each CM-step, the iteration is implemented for every generator once only, and is not repeated until convergence.

3. The Aitken δ^2 acceleration

The Aitken δ^2 method is a nonlinear method for accelerating the convergence of scalar sequences and it is particularly powerful for linear sequence convergence. The method is a simple and computationally inexpensive procedure.

First, we describe the Aitken δ^2 method. Let $\{\phi^{(r)}\}_{r\geq 0}$ be a scalar sequence which converges to ϕ^*. For the scalar sequence $\{\phi^{(r)}\}_{r\geq 0}$, the Aitken δ^2 method generates a sequence $\{\dot{\phi}^{(r)}\}_{r\geq 0}$ by

$$\dot{\phi}^{(r)} = \phi^{(r)} - \frac{(\phi^{(r+1)} - \phi^{(r)})^2}{\phi^{(r+2)} - 2\phi^{(r+1)} + \phi^{(r)}}. \tag{5}$$

For convergence of the scalar sequence $\{\dot{\phi}^{(r)}\}_{r\geq 0}$, Traub (1964) provided the following lemma.

Lemma 1. *If $\{\phi^{(r)}\}_{r\geq 0}$ converges to a stationary point ϕ^* as $r \to \infty$, then $\{\dot{\phi}^{(r)}\}_{r\geq 0}$ generated by Eq. (5) converges to the same stationary point ϕ^*.*

To compare the speed of convergence of the sequence $\{\dot{\phi}^{(r)}\}_{r\geq 0}$ from the Aitken δ^2 method with that of the sequence $\{\phi^{(r)}\}_{r\geq 0}$, we use the following notion of Brezinski and Zaglia (1991).

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 ϕ^* *faster than* $\{\phi^{(r)}\}_{r\geq 0}$ or that the extrapolation method accelerates the convergence of $\{\phi^{(r)}\}_{r\geq 0}$.

Traub (1964) proved that the Aitken δ^2 method accelerates the convergence of $\{\phi^{(r)}\}_{r\geq 0}$ in the sense that

$$\lim_{r\to\infty} \frac{|\dot{\phi}^{(r)} - \phi^*|}{|\phi^{(r+2)} - \phi^*|} = 0. \tag{6}$$

Next we apply the Aitken δ^2 acceleration to the EM and ECM algorithms. For simplicity, we write $\theta = \{p_V(x_V) | x_V \in \Omega_V\} = (\theta_1, \theta_2, \ldots, \theta_d)$, where d is the number of cells in the contingency table and the probability of the last cell is $\theta_d = 1 - \sum_{j=1}^{d-1} \theta_j$. Since the Aitken δ^2 method is an accelerator for a scalar sequence but not for a vector sequence, we transform the parameter vector θ into distinct scalar parameters. A multinomial distribution with d cells can be factorized into $d - 1$ independently conditional binomial distributions whose parameters $\varphi = \varphi(\theta)$ are defined respectively as

$$\begin{aligned} \varphi &= (\phi_1, \phi_2, \ldots, \phi_{d-1}) \\ &= \left(\theta_1, \frac{\theta_2}{1-\theta_1}, \ldots, \frac{\theta_{d-1}}{1 - (\sum_{j=1}^{d-2} \theta_j)} \right), \end{aligned} \tag{7}$$

(see the Appendix). Since $\varphi = \varphi(\theta)$ is a one–one function of θ, the stationary point of $\varphi = \varphi(\theta)$ is $\varphi(\theta^*)$ for the stationary point θ^* of θ. In order to accelerate the convergence of the sequence $\{\theta^{(r)}\}_{r\geq 0}$ obtained by the EM or ECM algorithms, we apply the Aitken δ^2 acceleration to generate the sequence $\{\dot{\theta}^{(r)}\}_{r\geq 0}$. The Aitken δ^2 acceleration for the EM and ECM algorithms is presented as follows.
Let $\theta^{(0)} = \dot{\theta}^{(0)}$ denote the initial value.

E-step: Using $\theta^{(r)} = \{p(x_V)^{(r)} | x_V \in \Omega_V\}$ and the observed frequencies, calculate the expected marginal counts $\tilde{n}_e^{(r)}$ for each generator $e \in E$ by Eq. (2).
M (or CM)-step: Find $\theta^{(r+1)} = \{p(x_V)^{(r+1)} | x_V \in \Omega_V\}$ by using Eq. (3) or (4).
Aitken δ^2acceleration: Calculate

$$\varphi^{(r-1)} = \varphi(\theta^{(r-1)}), \qquad \varphi^{(r)} = \varphi(\theta^{(r)}), \qquad \varphi^{(r+1)} = \varphi(\theta^{(r+1)})$$

by Eq. (7) where $(\theta^{(r-1)}, \theta^{(r)}, \theta^{(r+1)})$ is obtained at the previous M (or CM)-steps. Generate a vector $\dot{\varphi}^{(r-1)} = (\dot{\phi}_i^{(r-1)})_{i=1,\dots,d-1}$ from

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

Calculate $\dot{\theta}^{(r-1)}$ from

$$\begin{aligned}\dot{\theta}^{(r-1)} &= (\dot{\theta}_1^{(r-1)}, \dot{\theta}_2^{(r-1)}, \dots, \dot{\theta}_{d-1}^{(r-1)}, \dot{\theta}_d^{(r-1)}) \\ &= \left(\dot{\phi}_1^{(r-1)}, \dot{\phi}_2^{(r-1)}(1 - \dot{\phi}_1^{(r-1)}), \dots, \dot{\phi}_{d-1}^{(r-1)} \prod_{j=1}^{d-2}(1 - \dot{\phi}_j^{(r-1)}), 1 - \left(\sum_{j=1}^{d-1} \dot{\theta}_j^{(r-1)}\right)\right)\end{aligned} \tag{8}$$

and check the convergence by

$$\max_{1\leq i\leq d} |\dot{\theta}_i^{(r-1)} - \dot{\theta}_i^{(r-2)}| \leq \delta,$$

where δ is the desired accuracy.

Note that to find the expected frequencies at each E-step, we use $\theta^{(r)} = \{p(x_V)^{(r)} | x_V \in \Omega_V\}$ obtained at the previous M-step but not the $\dot{\theta}^{(r)}$ obtained at the previous Aitken δ^2 acceleration. Below we give the properties of the convergence of the Aitken δ^2 acceleration for the EM and ECM algorithms. Let θ^* denote a stationary point of the sequence $\{\theta^{(r)}\}_{r\geq 0}$ obtained by the EM or ECM algorithm.

Theorem 1. *For a given initial value $\theta^{(0)}$, the sequence $\{\dot{\theta}^{(r)}\}_{r\geq 0}$ obtained by using the Aitken δ^2 acceleration converges to the same stationary point θ^*.*

Proof. Since $\varphi = \varphi(\theta)$ of Eq. (7) is a one–one monotone function of θ, we need only prove convergence of the sequence $\{\dot{\varphi}^{(r)}\}_{r\geq 0}$ to the stationary point φ^* of the sequence $\{\varphi^{(r)}\}_{r\geq 0}$.

If $\{\varphi^{(r)}\}_{r\geq 0}$ is a convergent vector sequence to the vector φ^*, 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 $\|\cdot\|_\infty = \max_i\{|\cdot|\}$. Thus, from the above facts and Lemma 1, the sequence $\{\dot{\varphi}^{(r)}\}_{r\geq 0}$ converges to the stationary point φ^* of the sequence $\{\varphi^{(r)}\}_{r\geq 0}$ as $r \to \infty$. □

In the following theorem, we show the speed of convergence of the Aitken δ^2 accelerated EM and ECM algorithms.

Theorem 2. *The Aitken δ^2 acceleration speeds up the convergence of the EM and ECM algorithms, that is, for all i,*

$$\lim_{r\to\infty} \frac{|\dot{\theta}_i^{(r)} - \theta_i^*|}{|\theta_i^{(r+2)} - \theta_i^*|} = 0.$$

Proof. From Eq. (6), we have

$$\lim_{r\to\infty} \frac{|\dot{\phi}_i^{(r)} - \phi_i^*|}{|\phi_i^{(r+2)} - \phi_i^*|} = 0,$$

Table 1
A three-way contingency table with partially classified frequencies

Clinic (X_1)	Prenatal Care (X_2)	Survival (X_3)	
		Died	Survived
(a) *Completely classified cases*			
A	Less	3	176
	More	4	293
B	Less	17	197
	More	2	23
(b) *Partially classified cases (Clinic missing)*			
	Less	50	500
	More	25	150
(c) *Partially classified cases (Prenatal Care missing)*			
A		10	900
B		20	500

Source: (a) Bishop et al. (1975), Table 2.4-2. (b) and (c) Artificial data.

for all i. Since $\dot{\theta}^{(r)}$ of Eq. (8) is a one–one monotone function of $\dot{\varphi}^{(r)}$, it follows that

$$\lim_{r\to\infty} \frac{|\dot{\theta}_i^{(r)} - \theta_i^*|}{|\theta_i^{(r+2)} - \theta_i^*|} = 0,$$

for all i. Thus the sequence $\{\dot{\theta}^{(r)}\}_{r\geq 0}$ converges to θ^* faster than $\{\theta^{(r)}\}_{r\geq 0}$ does. □

4. Numerical experiments

In this section, we illustrate how much faster the Aitken δ^2 accelerated EM algorithm converges compared to the EM algorithm using numerical experiments. Note that the convergence speed of the Aitken δ^2 acceleration does not depend on the structure of the log-linear models but on the EM and ECM sequences.

Example 1. Consider a $2 \times 2 \times 2$ contingency table concerning the survival of infants in Table 1. The completely classified data in Table 1(a) was previously analyzed in Bishop et al. (1975), and Table 1(b) and (c) give the artificial data. For the data of Table 1(b), Clinic (X_1) is missing; for the data of Table 1(c), Prenatal Care (X_2) is missing. The observed pattern is $T = \{\{1, 2, 3\}, \{1, 2\}, \{2, 3\}\}$. Suppose that the data have a multinomial distribution with unknown parameter θ.

Table 2 summarizes the MLEs for several log-linear models and the numbers of iterations for the EM and ECM algorithms and the Aitken δ^2 acceleration with the desired accuracy $\delta = 10^{-9}$. We applied the ECM algorithm to the no third-order-interaction term model with $E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$. As shown in Table 2, the Aitken δ^2 acceleration converges to the MLEs faster than the EM algorithm, and the convergence speed is about twice as fast as that of the EM algorithm for all of these log-linear models.

Example 2. Consider a 2×2 contingency table with completely and partially classified observations. Let X_1 and X_2 be dichotomous variables. Suppose that the observed frequencies have a multinomial distribution with a parameter vector θ. The observed frequencies classified by X_1 and X_2 are

$$n_{12} = (n_{12}(1, 1), n_{12}(2, 1), n_{12}(1, 2), n_{12}(2, 2)) = (5, 2, 4, 1),$$

and the frequencies partially classified by X_1 are

$$n_1 = (n_1(1), n_1(2)) = (75, 25).$$

Several data sets of the observed frequencies n_2 partially classified by X_2 are given in the first two columns of Table 3. Set the desired accuracy δ to be 10^{-9}. From the third column of Table 3, we can see that for these data sets, the convergence of the EM algorithm is quite slow, and its convergence depends largely on the proportion of incompletely observed frequencies. From the fourth column of Table 3, we can see that Aitken δ^2 acceleration speeds up the convergence of the EM algorithm by a factor between 2 and 4.

Table 2
MLEs for each log-linear model and the numbers of iterations required by the EM and ECM algorithms and the Aitken δ^2 acceleration

Clinic (X_1)	Prenatal Care (X_2)	Survival (X_3)		Number of iterations	
		Died	Survived	EM or ECM	Aitken δ^2 acceleration
(a) Saturated model with $E = \{\{1, 2, 3\}\}$					
A	Less	0.0046	0.1450	137	61
	More	0.0140	0.3327		
B	Less	0.0133	0.4142		
	More	0.0018	0.0745		
(b) No three-interaction model with $E = \{\{1, 2\}, \{1, 3\}, \{2, 3\}\}$					
A	Less	0.0056	0.1438	133	59
	More	0.0129	0.3341		
B	Less	0.0125	0.4160		
	More	0.0027	0.0722		
(c) Conditional independence model with $E = \{\{1, 2\}, \{1, 3\}\}$					
A	Less	0.0057	0.1439	130	57
	More	0.0131	0.3338		
B	Less	0.0128	0.4151		
	More	0.0023	0.0722		
(d) Conditional independence model with $E = \{\{1, 2\}, \{2, 3\}\}$					
A	Less	0.0046	0.1446	131	57
	More	0.0129	0.3341		
B	Less	0.0131	0.4160		
	More	0.0028	0.0721		
(e) Conditional independence model with $E = \{\{1, 3\}, \{2, 3\}\}$					
A	Less	0.0010	0.2430	38	18
	More	0.0076	0.2244		
B	Less	0.0090	0.2595		
	More	0.0068	0.2396		

Table 3
The numbers of iterations required by the EM and the Aitken δ^2 acceleration

Total	n_2	Number of iterations	
		EM	Aitken δ^2 acceleration
200	(94, 106)	284	73
400	(233, 167)	42	12
600	(272, 328)	609	158
800	(471, 329)	261	116
1000	(467, 533)	898	205
1200	(679, 521)	605	178
1400	(654, 746)	1,182	262
1600	(704, 896)	1,351	342
1800	(900, 900)	1364	286
2000	(1012, 988)	1462	311
2200	(1144, 1056)	1499	333
2400	(1031, 1369)	1908	486
2600	(1440, 1160)	1338	368
2800	(1141, 1659)	2172	597
3000	(1410, 1590)	2234	479

Acknowledgements

The authors are grateful to an anonymous referee and the executive editor Dr. Richard A. Johnson for valuable comments. This research was supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Young Scientists, No. 187002796.

Appendix

Let $Y = (Y_1, \dots, Y_d)$ be a random vector from a multinomial distribution with a parameter vector $\theta = (\theta_1, \theta_2, \dots, \theta_d)$, where $\theta_d = 1 - (\sum_{i=1}^{d-1} \theta_i)$. That is, the distribution of Y is given by

$$f(y|\theta) = \frac{n!}{y_1! y_2! \cdots y_d!} \theta_1^{y_1} \theta_2^{y_2} \cdots \theta_d^{y_d},$$

where $n = \sum_{i=1}^{d} y_i$. Using a chain of binomial random variables, the multinomial distribution can be factorized into a product of conditional binomial distributions of Y_i's as follows:

$$\begin{aligned} f(y|\theta) &= \left[\frac{n!}{y_1!(n-y_1)!} \theta_1^{y_1} (1-\theta_1)^{n-y_1} \right] \\ &\quad \times \left[\frac{(n-y_1)!}{y_2!(n-y_1-y_2)!} \left(\frac{\theta_2}{1-\theta_1} \right)^{y_2} \left(\frac{1-\theta_1-\theta_2}{1-\theta_1} \right)^{n-y_1-y_2} \right] \\ &\quad \times \cdots \times \left[\frac{(n - (\sum\limits_{j=1}^{d-2} y_j))!}{y_{d-1}! y_d!} \left(\frac{\theta_{d-1}}{1 - (\sum\limits_{j=1}^{d-2} \theta_j)} \right)^{y_{d-1}} \left(\frac{\theta_d}{1 - (\sum\limits_{j=1}^{d-2} \theta_j)} \right)^{y_d} \right]. \end{aligned}$$

Thus we can obtain the parameterization $\varphi = \varphi(\theta)$ given by

$$\begin{aligned} \varphi &= (\phi_1, \phi_2, \dots, \phi_{d-1}) \\ &= \left(\theta_1, \frac{\theta_2}{1-\theta_1}, \dots, \frac{\theta_{d-1}}{1 - (\sum\limits_{j=1}^{d-2} \theta_j)} \right). \end{aligned}$$

References

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

Bishop, Y.M.M., Fienberg, S.E., Holland, P.W., 1975. Discrete Multivariate Analysis: Theory and Practice. The M.I.T. Press, Cambridge, MA, London.

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

Chen, T.T., Hochberg, Y., Tenenbein, A., 1984. Analysis of multivariate categorical data with misclassification errors by triple sampling schemes. J. Statist. Planing and Inference 9, 177–184.

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.

Fuchs, C., 1982. Maximum likelihood estimation and model selection in contingency tables with missing data. J. Amer. Stat. Assoc. 77, 270–278.

Lauritzen, S.L., 1995. The EM algorithm for graphical association models with missing data. Comput. Statist. Data Anal. 19, 191–201.

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

Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80, 267–278.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581–592.

Traub, J.F., 1964. Iterative Methods for the Solution of Equations. Prentice-Hall, Inc., Englewood Cliffs, NJ.