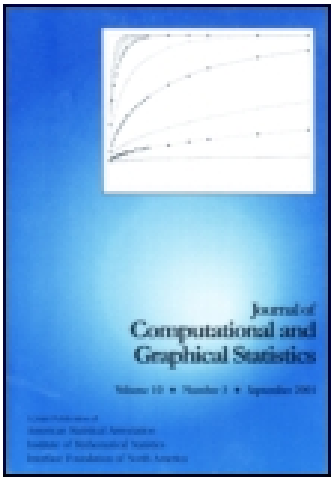

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

A Simple Method for Computing the Observed Information Matrix When Using the EM Algorithm with Categorical Data

Stuart G. Baker [a]

[a] Mathematical Statistician, Screening Section, Biometry Branch, Division of Cancer Prevention and Control , National Cancer Institute , EPN 344, 9000 Rockville Pike, Bethesda , MD , 20892 , USA

Published online: 21 Feb 2012.

To cite this article: Stuart G. Baker (1992) A Simple Method for Computing the Observed Information Matrix When Using the EM Algorithm with Categorical Data, Journal of Computational and Graphical Statistics, 1:1, 63-76

To link to this article: http://dx.doi.org/10.1080/10618600.1992.10474576

A Simple Method for Computing the Observed Information Matrix When Using the EM Algorithm With Categorical Data

STUART G. BAKER*

A simple matrix formula is given for the observed information matrix when the EM algorithm is applied to categorical data with missing values. The formula requires only the design matrices, a matrix linking the complete and incomplete data, and a few simple derivatives. It can be easily programmed using a computer language with operators for matrix multiplication, element-by-element multiplication and division, matrix concatenation, and creation of diagonal and block diagonal arrays. The formula is applicable whenever the incomplete data can be expressed as a linear function of the complete data, such as when the observed counts represent the sum of latent classes, a supplemental margin, or the number censored. In addition, the formula applies to a wide variety of models for categorical data, including those with linear, logistic, and log-linear components. Examples include a linear model for genetics, a log-linear model for two variables and nonignorable nonresponse, the product of a log-linear model for two variables and a logit model for nonignorable nonresponse, a latent class model for the results of two diagnostic tests, and a product of linear models under double sampling.

Key Words: Incomplete data; Missing data; Poisson distribution.

1. INTRODUCTION

The EM algorithm (Dempster, Laird, and Rubin 1977) is a convenient method for finding maximum likelihood (ML) estimates when data are missing. According to Redner and Walker (1984) "its [the EM algorithm's] most appealing general property is that it produces sequences of iterates on which the log-likelihood function increases monotonically ... [and] ... does not require augmentation with elaborate safeguards, such as those necessary for Newton's method and quasi-Newton methods, in order to produce iteration sequences with good global convergence characteristics." Another desirable property is the ease of implementation, particularly when data are categorical. When sufficient statistics are sums of cell counts, as is often the case with categorical data, the E step simply

*Stuart G. Baker is Mathematical Statistician, Screening Section, Biometry Branch, Division of Cancer Prevention and Control, National Cancer Institute, EPN 344, 9000 Rockville Pike, Bethesda, MD 20892

Received June 1991; Revised January 1992

Volume 1, Number 1, pp. 63–76

imputes cell counts. The *M* step uses standard techniques to maximize a likelihood for the imputed counts.

One criticism of the EM algorithm is that it does not routinely compute the observed information matrix, which is important for Bayesian and likelihood-based inference. Although various methods have been proposed for computing the observed information matrix when applying the EM algorithm, they are difficult to implement with many common models for categorical data. These difficulties include numerical inaccuracies, intensive analytical work, complicated computer programming, and large amounts of central processing unit (CPU) time.

The proposed method avoids the aforementioned difficulties because it is a simple matrix formula. Consequently it is easy to program using any computer language with matrix operators. It runs quickly since it involves no iterative procedures. The only analytical work involves specification of a few simple derivatives. Because the formula is an algebraic expression rather than a numerical approximation, numerical inaccuracies are minimized.

This method is applicable to a wide variety of problems in which the EM algorithm is used to maximize a likelihood involving missing categorical data. More specifically, it can be applied whenever the counts for the incomplete data can be expressed as a linear function of the counts for the complete data. (The terms *complete data* and *incomplete data* are defined in Dempster, Laird, and Rubin [1977].) Applicable problems include those in which observed counts are the sum of counts for latent classes, form a supplemental margin, or equal the number censored. Right censoring is easily included in this framework by defining a category for survival past the longest time in the study.

2. REVIEW OF THE LITERATURE

Methods for estimating standard errors when applying the EM algorithm are based on either the observed information matrix, the expected information matrix, or on resampling techniques. These methods are reviewed in the context of categorical data.

2.1 THE OBSERVED INFORMATION MATRIX

Previous methods for computing the observed information matrix are applicable to both continuous and categorical data. The generality does not allow them to exploit the special structure of incomplete categorical data.

The earliest method for computing the observed information matrix in the context of the EM algorithm was proposed by Hartley (1958) and Hartley and Hocking (1971). (At the time the general structure of the EM algorithm was not known, so the method was applied to special cases of the EM.) The observed information matrix is computed by creating a system of simultaneous equations, one for each EM iteration, based on the estimates and the incomplete-data-score vectors. Computation of the incomplete-data-score vector is easily accomplished by substituting the imputed cell counts into the complete-data-score vector. Thus the method can be implemented for simple models. Unfortunately, it cannot be applied to complicated models for which the number of

parameters is greater than the number of equations (iterations).

The method of EM-aided differentiation (Meilijson 1989) computes the observed information matrix by perturbing the score vector for the incomplete data. A caveat is that inaccuracies may arise when numerical differentiation procedures are applied to sparse data.

The method of Meng and Rubin (1991) mitigates the potential for numerical inaccuracies by using only numerical techniques to compute the derivative of the function mapping the estimate from one iteration to the next. A limitation is that it requires code for the complete-data-observed information matrix, which is not available for some complicated models.

Unlike the previously described methods, that of Louis (1982) computes the observed information matrix without using numerical approximations. Instead, algebraic expectations, conditional on the incomplete data, are taken for the complete-data-information matrix, the complete-data-score vector, and the square of the complete-data-score vector. For some problems, such as mixtures, the algebra is manageable, but for other problems it is difficult, particularly if the complete-data-information matrix also has to be derived. Unfortunately, the simple formula given for the multinomial model applies only to those models in which the second derivative of the expected cell counts equals 0.

Wei and Tanner (1990) describe a modification of Louis (1982) that substitutes Monte Carlo techniques for algebraic integration. This method is described in the context of an entire Monte Carlo implementation of the EM algorithm but could be used at the completion of the usual EM algorithm. Like some of the other methods, a drawback is that it requires code for the complete-data-information matrix, which may not be available.

Ibrahim (1990) describes an application of Louis (1982) to generalized linear models that yields a matrix formula for the observed information matrix. However, use of this formula is limited to problems in which only the covariates, and not the response variable, are missing.

2.2 THE EXPECTED INFORMATION MATRIX

In the case of categorical data with missing values, the expected information matrix is generally easier to compute than the observed information matrix. Although the methods for obtaining the expected information matrix were proposed as part of Fisher scoring algorithms, these methods can also be invoked at the completion of the EM algorithm.

The easiest of these methods to implement is the method of composite link functions for generalized linear models (Thompson and Baker 1981). The composite link function is a matrix mapping the complete data to the incomplete data. Unfortunately, the method applies only to a limited class of models.

To accommodate somewhat more complicated models, Thompson and Baker (1981) proposed bilinear link functions for generalized linear models, which Burn (1983) applied to a model for misclassification. More analytical work is required to compute the expected information matrix, however, and the method does not apply to all models of interest.

Espeland (1986) extended the idea of composite link function to linear, product, and

log-linear models. For these models the expected information can be easily computed using matrices defined in the model. Unfortunately, the method does not apply to some categorical data models of interest—recursive systems of log-linear models and models for discrete time hazard functions.

In contrast, methods for computing the expected information matrix based on nonlinear regression (Green 1984; Jorgensen 1983; and Palmgren and Ekholm 1987) are applicable to any model. The generality comes at a price, however—the user must algebraically or numerically differentiate the expected counts for the incomplete data.

Meilijson (1989) suggested using the empirical information matrix, which is a consistent estimator of the expected information matrix. The empirical information is relatively easy to compute since it requires only the incomplete-data-score vector for each iid component of the likelihood. Using the method of Hartley (1958), the incomplete-data-score vector can be easily obtained from the complete-data-score vector. A drawback to the use of the empirical information matrix is that it is inefficient and ignores the likelihood principle.

2.3 RESAMPLING METHODS

Resampling methods have also been used to obtain standard errors with the EM algorithm. The bootstrap approach (Efron 1982) can be readily implemented with categorical data (e.g., Baker and Laird 1988 and Baker and Chu 1990). A drawback is the large amounts of CPU time that are often required.

Jorgensen (1987) devised a jackknife procedure that substantially reduces the CPU time. Instead of performing the entire EM algorithm for each jackknife iteration, Jorgensen's method performs a one-step iteration of the EM algorithm away from convergence. The main drawback is the need to algebraically differentiate the function mapping the parameter values from one iteration to the next. Alternatively the method in Meng and Rubin (1991) might be used to numerically obtain this derivative. Unfortunately programming both the one-step iteration and the numerical derivative may be difficult.

3. THE PROPOSED METHOD

For many problems it is desirable to compute standard errors based on the observed information matrix rather than the expected information matrix or resampling. Efron and Hinkley (1978) have argued that the observed information matrix is preferable to the expected information for conditional inference. For small-to-moderate samples with complicated patterns of missing data, the performance of resampling techniques is unclear. Moreover, for likelihood-based or Bayesian inference, the observed information matrix is the desired quantity (Meng and Rubin 1991).

For computational reasons investigators desiring standard errors based on the observed information matrix often had to settle for standard errors computed using the expected information matrix or resampling techniques. Using the proposed method when applying the EM algorithm to categorical data, investigators can easily compute the observed information matrix.

The proposed method requires the use of special matrix functions. The notation for these functions follows:

matrix1 · matrix2 ≡ element-by-element multiplication of matrix1 and matrix2,
matrix1/matrix2 ≡ element-by-element division of matrix1 by matrix2,
$\mathbf{1}_{n\times 1}$ ≡ an $n \times 1$ matrix of 1's,
diag(vector) ≡ a matrix with vector on the diagonal and 0's otherwise,
block(matrix1, matrix2) ≡ a matrix with matrix1 and matrix2 forming a block diagonal, with 0's otherwise,
hcat(matrix1, matrix2) ≡ horizontal concatenation of matrix1 and matrix2,
vcat(matrix1, matrix2) ≡ vertical concatenation of matrix1 and matrix2.

The functions *block*( ), *hcat*( ), and *vcat*( ) can have more than two arguments and can be applied to indexed arguments, as in $block_{|_k}$ matrix(k).

3.1 PRELUDE TO THE MATRIX FORMULA: A MATRIX EM ALGORITHM

The proposed method for computing the observed information matrix is particularly easy to implement when a matrix formulation has been used for the EM algorithm with categorical data. This section describes such a matrix formulation.

Assume the categorical complete data follow a Poisson, multinomial, or product multinomial distribution, with a parameter vector denoted by $\boldsymbol{\theta}_{p\times 1}$. Let $i = 1, 2, \ldots, m$ index the cells for the categorical complete data, and let $j = 1, 2, \ldots, m$ index the cells for the categorical incomplete data, where $n \geq m$. I use y_i^* to denote the count for the ith cell of the complete data and y_j to denote the count for the jth cell of the incomplete data. Similarly, I use $\mu_j(\boldsymbol{\theta}) = E(y_j)$ and $\mu_i^*(\boldsymbol{\theta}) = E(y_i^*)$ to denote the corresponding expected counts. Each expected count for the complete data equals the sum of certain expected counts for the incomplete data. More formally, I relate these expected counts by the following equation: $\mu_j \equiv \sum_i c_{ji}\mu_i^*$, where c_{ji} equals 0 or 1.

To introduce the matrix notation I define the following quantities. Let $\mathbf{Y}^*_{n\times 1} = [y_1^* \, y_2^* \ldots y_i^* \ldots y_n^*]'$ and $\mathbf{Y}_{m\times 1} = [y_1 \, y_2 \ldots y_j \ldots y_m]'$ denote the complete and incomplete data. Let $\mathbf{U}^*_{n\times 1}(\boldsymbol{\theta}) = E(\mathbf{Y}^*) = [\mu_1^*(\boldsymbol{\theta}) \ldots \mu_i^*(\boldsymbol{\theta}) \ldots \mu_n^*(\boldsymbol{\theta})]'$ and $\mathbf{U}_{m\times 1}(\boldsymbol{\theta}) = E(\mathbf{Y}) = [\mu_1(\boldsymbol{\theta}) \ldots \mu_j(\boldsymbol{\theta}) \ldots \mu_m(\boldsymbol{\theta})]'$ denote the expected cell counts for the complete and incomplete data. Following Thompson and Baker (1981) and Espeland and Odoroff (1985), I relate $\mathbf{U}_{m\times 1}(\boldsymbol{\theta})$ and $\mathbf{U}^*_{n\times 1}(\boldsymbol{\theta})$ by the following equation: $\mathbf{U}_{m\times 1}(\boldsymbol{\theta}) = \mathbf{C}_{m\times n}\mathbf{U}^*_{n\times 1}(\boldsymbol{\theta})$, where $\mathbf{C}$ has elements $\{c_{ji}\}$.

The matrix form for the EM algorithm with categorical data (Espeland and Odoroff 1985) is:

$$
\begin{aligned}
&E \text{ Step: } \mathbf{Y}^*(\boldsymbol{\theta}^{(t)}) = \mathbf{U}^*(\boldsymbol{\theta}^{(t)}) \cdot \mathbf{C}'(\mathbf{Y}/(\mathbf{C}\mathbf{U}^*(\boldsymbol{\theta}^{(t)}))) \\
&M \text{ Step: Compute } \mathbf{U}^*(\boldsymbol{\theta}^{(t+1)}), \\
&\qquad \text{where } \boldsymbol{\theta}^{(t+1)} \text{ is the ML estimate based on } \mathbf{Y}^*(\boldsymbol{\theta}^{(t)}).
\end{aligned} \tag{3.1}
$$

3.2 DERIVATION OF THE MATRIX FORMULA

In this section I derive the simple matrix formula for the observed information matrix. The general strategy is to expand on the use of the composite link matrix for computing the expected information matrix.

To simplify the derivation I compute the observed information matrix assuming the complete data follow a Poisson distribution. When the complete data follow a multinomial or product multinomial distribution, there are two cases to consider. I discuss each case with respect to a multinomial distribution, but the discussion can be readily generalized to a product multinomial distribution. Suppose $\{y_1, \ldots, y_i, \ldots, y_n\} \sim \text{mult}(\mu^*_{Mi}(\boldsymbol{\theta}_M)/\sum_{i=1}^m \mu^*_{Mi}(\boldsymbol{\theta}_M), i = 1, 2, \ldots, n; N)$, where N is a constant, and $\boldsymbol{\theta}_M$ is a vector of parameters. To compute the observed information matrix, I assume $\{y_1, \ldots, y_i, \ldots, y_n\} \sim \text{Poisson}(\mu_i(\boldsymbol{\theta}), i = 1, 2, \ldots, n)$, where the relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_M$, and between $\mu_i(\boldsymbol{\theta})$ and $\mu^*_{Mi}(\boldsymbol{\theta}_M)$, depends on the case.

Case 1: The parameterization constrains $\sum_{j=1}^m \mu_{Mj}(\boldsymbol{\theta}_M)$ to equal N, where $\mu_{Mj} \times(\boldsymbol{\theta}_M) = \sum_{i=1}^n c_{ji}\mu^*_{Mi}(\boldsymbol{\theta}_M)$. In this case, by setting $\boldsymbol{\theta} = \boldsymbol{\theta}_M$ and $\mu^*_i(\boldsymbol{\theta}) = \mu^*_{Mi}(\boldsymbol{\theta}_M)$, the kernels of the Poisson and multinomial log-likelihoods are identical, namely, $\ell(\boldsymbol{\theta}) = \sum_{j=1}^m y_j \log(\mu_j(\boldsymbol{\theta}))$. Therefore, the observed information matrix derived from the Poisson log-likelihood equals that derived from the multinomial log-likelihood. (This case is also satisfied when $\sum_{i=1}^n \mu^*_{Mi}(\boldsymbol{\theta}_M) = N$ and $\sum_{j=1}^m c_{ji} = 1$, for all i.)

Case 2: The parameterization does not constrain $\sum_{j=1}^m \mu_{Mj}(\boldsymbol{\theta}_M)$ to equal N. In this case it is necessary to include a nuisance parameter ϕ in the following manner: $\mu^*_i(\boldsymbol{\theta}) = \exp(\phi)\mu^*_{Mi}(\boldsymbol{\theta}_M)$, where $\boldsymbol{\theta} = vcat(\phi, \boldsymbol{\theta}_M)$. The observed information matrix derived from the Poisson log-likelihood with the nuisance parameter yields the correct asymptotic variance–covariance matrix for $\widehat{\boldsymbol{\theta}}_M$ (Baker 1992; Palmgren 1981).

To simplify the notation I suppress the bounds of summation ($j = 1, 2, \ldots, m$; $i = 1, 2, \ldots, n$) and drop the argument θ. Under the Poisson distribution the kernel of the log-likelihood for the incomplete data is $\ell = \sum_j y_j \log(\sum_i c_{ji}\mu^*_i) - \sum_j \sum_i c_{ji}\mu^*_i$, and the first derivative is $\dot{\ell} = \sum_j (\sum_i c_{ji}\, \dot{\mu}^{*\prime}_i)(y_j/\mu_j - 1)$. The negative of the second derivative (the observed information matrix when evaluated at the maximum likelihood estimate) is

$$\mathcal{I} = \sum_j \sum_i c_{ji}\, \ddot{\mu}^*_i(1 - y_j/\mu_j) + \sum_j \left(\sum_i c_{ji}\, \dot{\mu}^*_i\right)' \left(\sum_i c_{ji}\, \dot{\mu}^*_i\right) y_j/\mu_j^2. \tag{3.2}$$

By formulating (3.2) in matrix notation, I can simplify evaluation of the observed information matrix, as the formula can be easily coded in matrix programming languages such as SAS IML, APL, or GAUSS.

To write (3.2) in matrix notation, I introduce the following general class of models for the expected counts for the complete data:

$$\mathbf{U}^* \propto \exp\left[\sum_{k=1}^K \mathbf{G}^{(k)}\mathbf{H}^{(k)}\right], \tag{3.3}$$

where $\mathbf{H}^{(k)} = h^{(k)}(\mathbf{X}^{(k)}(\boldsymbol{\theta}^{(k)}; \mathbf{Z}^{(k)})$ is a $\tau \times 1$ matrix with elements $\{h_q^{(k)}\}$, $\boldsymbol{\theta}^{(k)}$ denotes

a $\rho^{(k)} \times 1$ subset of parameters $\left(\sum_k \rho^{(k)} = \rho\right)$, $\mathbf{X}^{(k)}$ is a $\tau \times \rho^{(k)}$ design matrix with rows $\{x_q^{(k)}\}$ and elements $\{x_{qa}^{(k)}\}$, $\mathbf{Z}^{(k)}$ is a $\tau \times 1$ matrix with elements $\{z_q^{(k)}\}$, and $\mathbf{G}^{(k)}$ is an $n \times \tau$ matrix with elements $\{g_{iq}^{(k)}\}$. Also $h^{(k)}(\ )$ is a function that operates on each element of a vector, such as

$$h^{(k)}(\mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)}; Z^{(k)}) = \log(\mathbf{Z}^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)}),$$
$$h^{(k)}(\mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)}; Z^{(k)}) = \mathbf{Z}^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)},$$

and

$$h^{(k)}(\mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)}; Z^{(k)}) = (\mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)}) \cdot \mathbf{Z}^{(k)}) - \log(1 + \exp(\mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)})).$$

In many models $\mathbf{G}^{(k)}$ is the identity matrix. An example of $\mathbf{G}^{(k)}$ that is not the identity matrix is a matrix that maps discrete time hazards into probabilities (e.g., Baker, Wax, and Patterson 1992).

The derivation of the matrix formula requires a return to subscript notation. Rewriting (3.3) in subscript notation gives

$$u_i^* \propto \exp\left[\sum_k \omega_i^{(k)}\right], \tag{3.4}$$

where $\omega_i^{(k)} = \sum_q g_{iq}^{(k)} h_q^{(k)}$ and $h_q^{(k)} = h^{(k)}(x_q^{(k)}\boldsymbol{\theta}^{(k)}; z_q^{(k)})$. Let $\boldsymbol{\theta}_b$ denote the bth element in $\boldsymbol{\theta}_{\rho\times 1} = vcat(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(K)})$. Then

$$\dot{\mu}_{ib} \equiv \frac{\partial \mu_i}{\partial \boldsymbol{\theta}_b} = \mu_i^* \sum_k \dot{\omega}_{ib}^{(k)}, \tag{3.5}$$

$$\ddot{\mu}_{iab} \equiv \frac{\partial^2 \mu_i}{\partial \boldsymbol{\theta}_a \partial \boldsymbol{\theta}_b} = \mu_i^* \sum_k \ddot{\omega}_{iab}^{(k)} + \mu_i^* \sum_k \dot{\omega}_{ia}^{(k)} \sum_k \dot{\omega}_{ib}^{(k)}, \tag{3.6}$$

where $\dot{\omega}_{ib}^{(k)} = \delta_b^{(k)} \sum_q g_{iq}^{(k)} \dot{h}_q^{(k)} x_{qb}^{(k)}$, $\ddot{\omega}_{iab}^{(k)} = \delta_a^{(k)} \delta_b^{(k)} \sum_q g_{iq}^{(k)} \ddot{h}_q^{(k)} x_{qa}^{(k)} x_{qb}^{(k)}$, $\delta_b^{(k)} = 1$ if b indexes a parameter in $\boldsymbol{\theta}^{(k)}$, 0 otherwise, and $\dot{h}_q^{(k)}$ and $\ddot{h}_q^{(k)}$ denote the first and second derivatives of $h_q^{(k)}$ with respect to $(X_q^{(k)}\boldsymbol{\theta}^{(k)})$.

Define $r_i \equiv \sum_j c_{ji}(1 - y_j/\mu_j)$ and let c'_{ij} denote the (i, j) element of $\mathbf{C}'$. I write the (a, b) element of $\mathcal{I}$ as the sum of three terms:

$$\mathcal{I}_{ab} = \mathcal{I}_{1ab} + \mathcal{I}_{2ab} + \mathcal{I}_{3ab},$$

where

$$\mathcal{I}_{1ab} = \sum_j \sum_i c_{ji} \left[\mu_i^* \sum_k \ddot{\omega}_{iab}^{(k)}\right](1 - y_j/\mu_j)$$
$$= \sum_k \sum_l \sum_j c_{ji}(1 - y_j/\mu_j)\mu_i^* \ddot{\omega}_{iab}^{(k)} \qquad \text{(by rearranging terms)}$$

$$= \sum_k \sum_l r_i \mu_i^* \delta_a^{(k)} \delta_b^{(k)} \sum_q g_{iq}^{(k)} \ddot{h}_q^{(k)} x_{qa}^{(k)} x_{qb}^{(k)} \qquad \text{(by substitution)}$$

$$= \sum_k \delta_a^{(k)} \delta_b^{(k)} \sum_q \left(\ddot{h}_q^{(k)} \left(\sum_t g_{iq}^{(k)} (r_i \mu_i^*) \right) \right) x_{qa}^{(k)} x_{qb}^{(k)} \qquad \text{(by rearranging terms)}$$

$$\mathcal{I}_{2ab} = \sum_j \sum_i c_{ji} \left[\mu_i^* \sum_k \dot{\omega}_{ia}^{(k)} \sum_k \dot{\omega}_{ib}^{(k)} \right] (1 - y_j/\mu_j)$$

$$= \sum_t r_i \mu_i^* \left(\sum_k \dot{\omega}_{ia}^{(k)} \right) \left(\sum_k \dot{\omega}_{ib}^{(k)} \right) \qquad \text{(by substitution)}$$

$$\mathcal{I}_{3ab} = \sum_j \left[\left[\sum_i c_{ji} \dot{\mu}_{ia}^* \right] \left[\sum_i c_{ji} \dot{\mu}_{ib}^{(*)} \right] \right] y_j/\mu_j^2$$

$$= \sum_j \left[\sum_i c_{ji} \left(\sum_k \dot{\omega}_{ia}^{(k)} \dot{\mu}_i^{(*)} \right) \right] \left[\sum_i c_{ji} \left(\sum_k \dot{\omega}_{ib}^{(k)} \mu_i^* \right) \right] y_j/\mu_j^2 \qquad \text{(by substitution)}$$

$$= \sum_j (y_j/\mu_j^2) \left[\sum_i \mu_i^{(*)} c'_{ij} \sum_k \dot{\omega}_{ia}^{(k)} \right] \left[\sum_i \mu_i^{(*)} c'_{ij} \sum_k \dot{\omega}_{ib}^{(k)} \right]. \qquad (3.7)$$

To write (3.7) in matrix notation, I use the following relations:

$$\text{vector1}_d \, \text{vector2}_d = (\text{vector1} \cdot \text{vector2})_d,$$

$$\sum_d \text{matrix1}_{de} \, \text{vector1}_d = (\text{matrix1}' \, \text{vector1})_e,$$

and

$$\sum_d \text{vector1}_d \, \text{matrix1}_{de} \, \text{matrix2}_{df} = (\text{matrix1}' \, \text{diag[vector1]matrix2})_{ef},$$

where $\{\text{vector1}_d\}$, $\{\text{vector2}_d\}$, $\{\text{matrix}_{de}\}$, and $\{\text{matrix}_{df}\}$ denote elements of vector1, vector2, matrix1, and matrix2.

Let $\mathbf{S}$ be an $n \times \rho$ matrix with elements $\left\{ \sum_k \dot{\omega}_{ib}^{(k)} = \sum_k \delta_b^{(k)} \sum_q \dot{h}_q^{(k)} g_{qi}^{(k)\prime} x_{qb}^{(k)} \right\}$, where $g_{qi}^{(k)\prime}$ is the (q, i) element of $\mathbf{G}^{(k)\prime}$; let R be an $n \times 1$ matrix with elements $\{r_i\}$; let $\dot{\mathbf{H}}^{(k)}$ be an $n \times 1$ matrix with elements $\left\{ \dot{h}_q^{(k)} \right\}$; and let $\ddot{\mathbf{H}}^{(k)}$ be an $n \times 1$ matrix with elements $\left\{ \ddot{h}_q^{(k)} \right\}$. The matrix formula for the observed information matrix is

$$\mathcal{I} = \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3,$$

where

$$\mathcal{I}_1 = block|_k \mathbf{X}^{(k)\prime} diag \left(\ddot{\mathbf{H}}^{(k)} \cdot (\mathbf{G}^{(k)\prime} (\mathbf{R} \cdot \mathbf{U}^*)) \right) \mathbf{X}^{(k)},$$

$$\mathcal{I}_2 = \mathbf{S}' diag(\mathbf{R} \cdot \mathbf{U}^*) \mathbf{S},$$

$$\mathcal{I}_3 = \mathbf{S}' diag(\mathbf{U}^*) \mathbf{C}' diag(\mathbf{Y}/(\mathbf{U} \cdot \mathbf{U})) \mathbf{C} diag(\mathbf{U}^*) \mathbf{S},$$

and

$$\mathbf{R} = 1 - \mathbf{C}'(\mathbf{Y}/\mathbf{U}),$$
$$\mathbf{S} = hcat|_k \mathbf{G}^{(k)} diag\left(\dot{\mathbf{H}}^{(k)}\right) \mathbf{X}^{(k)}. \tag{3.8}$$

The expected information matrix is obtained by substituting $E(\mathbf{Y}) = \mathbf{U}$ into (3.8) and equals $\mathbf{S}'$ $diag(\mathbf{U}^*)$ $\mathbf{C}'$ $diag(\mathbf{1}/\mathbf{Y})$ $\mathbf{C}$ $diag(\mathbf{U}^*)$ $\mathbf{S}$. If the data fit perfectly, that is, $\mathbf{Y} = \mathbf{U}$, the observed and expected information matrices are identical.

3.3 IMPLEMENTATION OF THE MATRIX FORMULA

Implementation of the matrix formula is straightforward using any computer language with matrix functions. A subroutine can be written that can be inserted after the code for the EM algorithm. The required inputs are $\mathbf{Y}$, $\mathbf{U}^*$, $\mathbf{C}$, $\mathbf{X}^{(k)}$, $\dot{k}^{(k)}(\ )$, $\ddot{k}^{(k)}(\ )$, and $\mathbf{G}^{(k)}$, for $k = 1, 2, \ldots, K$. The quantities $\mathbf{Y}$, $\mathbf{C}$, and $\mathbf{X}^{(k)}$ are available from the EM algorithm. (It is easy to create the composite link matrix, $\mathbf{C}$, using functions for creating block diagonal matrices; it is also easy to create the design matrix, $\mathbf{X}^{(k)}$, using the Kronecker product function. See Section 4 for examples.) The vector of expected cell counts $\mathbf{U}^*$ is the output of the EM algorithm. The user must supply $\dot{k}^{(k)}(\ )$, $\ddot{k}^{(k)}(\ )$, and $\mathbf{G}^{(k)}$, which for most models, is easy to specify.

Ironically the most complicated aspect of the entire matrix EM procedure is the part usually regarded as the simplest, namely, the specification of the *M* step. That is because, unlike the specification of the *E* step and the computation of the observed information matrix, it varies considerably among models.

4. EXAMPLES

The method is illustrated using five examples in which the EM algorithm is applied to categorical data with missing values. Whenever contingency table data are reshaped as a vector, the first variable listed in the cross-classification varies fastest in the vector, the second varies second fastest, and so forth. Additional matrix notation is needed for the examples:

$$\begin{aligned} \text{matrix1} \otimes \text{matrix2} &\equiv \text{Kronecker product of matrix1 and matrix2,} \\ \mathbf{I}_n &\equiv n \times n \text{ identity matrix [vector1 = scalar1],} \\ &\equiv a \text{ vector with elements equal to 1, if the corresponding elements} \\ &\quad \text{in vector1 equal scalar1, and equal to 0 otherwise.} \end{aligned}$$

The calculations were programmed in SAS IML. A copy of the code is available from the author. In all the examples $\mathbf{G}^{(k)}$ is the identity matrix. For an example in which $\mathbf{G}^{(k)}$ is not the identity matrix, see Baker, Wax, and Patterson (1992).

The sampling distribution in all examples is multinomial or product multinomial. In Examples 4.1, 4.4, and 4.5, $\sum_j c_{ji} = 1$, and the model for the logarithm of the expected cell counts includes the term log(N). Therefore the kernels of the Poisson

and the multinomial or product multinomial log-likelihoods are identical, so no extra parameters are needed for using (3.8) to obtain the asymptotic variances. In Examples 4.2 and 4.3, the kernels are not identical. Therefore before using (3.8) to compute the asymptotic variances, I include the parameter ϕ in the model in the manner described in Section 3.2.

4.1 A LINEAR MODEL FOR GENETIC LINKAGE

This example was presented in Rao (1965) and discussed in Dempster, Laird, and Rubin (1977), Louis (1982), Jorgensen (1983), Tanner and Wong (1987), Meng and Rubin (1991), and Wei and Tanner (1990).

Random Variable. The random variable is the number in each genetic class.

Complete Data. The complete data are a 5×1 array of counts.

Incomplete Data. The incomplete data are $\mathbf{Y} = [125\ 18\ 20\ 34]'$. The first count is the sum of the first two counts for the complete data. The last three counts equal the last three counts for the complete data.

Composite Link. The composite link matrix is $\mathbf{C} = block([1\ 1], \mathbf{I}_3)$.

Model. The genetic linkage model is $\log(\mathbf{U}^*) = \log(N) + \log(\mathbf{Z}^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)})$, where $N = 197$, $\mathbf{Z}^{(1)} = [.5\ 0\ .25\ .25\ 0]'$, and $\mathbf{X}^{(1)} = [0\ .25\ -.25\ -.25\ .25]'$.

EM Algorithm. The E step is given in (3.1). The M step is $\widehat{\boldsymbol{\theta}} = (\mathbf{Y}^{*\prime}[(\mathbf{X}^{(1)}/.25) = 1])/(\mathbf{Y}^{*\prime}|(\mathbf{X}^{(1)}/.25)|)$. The algorithm converges to $\widehat{\boldsymbol{\theta}} = .62682$.

Observed Information Matrix. The observed information matrix is computed via (3.8) with $\mathbf{G}^{(1)} = \mathbf{I}_5$, $\dot{\mathbf{H}}^{(1)} = \mathbf{I}_{5\times 1}/(\mathbf{Z}^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)})$, and $\ddot{\mathbf{H}}^{(1)} = -\mathbf{1}_{5\times 1}/(\mathbf{Z}^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)})^2$. The standard error for $\widehat{\boldsymbol{\theta}}^{(1)}$, based on the observed information matrix, is .026123, which agrees with the value obtained by the aforementioned authors.

4.2 A LOG-LINEAR MODEL FOR SURVEY DATA

This example was discussed in Little and Rubin (1987), but no standard errors were provided.

Random Variables. For this discussion the random variables are called COV (covariate, at one of two levels), OUT (outcome, at one of two levels), and RI (response indicator, either RI = 1 if OUT is observed, or RI = 0, if OUT is missing).

Complete Data. The complete data are the $2 \times 2 \times 2$ cross-classification of COV, OUT, and RI, reshaped as an 8×1 vector.

Incomplete Data. The incomplete data for RI = 1 are $\mathbf{Y}_{RI=1} = [100\ 20\ 30\ 50]'$, which is a 2×2 cross-classification of COV and OUT, reshaped as 4×1 vector. The incomplete data for RI = 0 are $\mathbf{Y}_{RI=0} = [40\ 60]']$, which correspond to the two levels of COV, since OUT is not observed. Thus, all of the incomplete data are summarized by $\mathbf{Y} = vcat(Y_{RI=1}, \mathbf{Y}_{RI=0})$.

Composite Link. The composite link matrix is $\mathbf{C} = block(\mathbf{I}_4, (\mathbf{I}_2 \otimes \mathbf{1}_{1\times 2}))$.

Model. A log-linear model for the two variables and nonignorable nonresponse is $\log(\mathbf{U}^*) = \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)}$, where $\boldsymbol{\theta}^{(1)} = [\phi, \theta^{(1)}_{\text{COV}}, \theta^{(1)}_{\text{OUT}}, \theta^{(1)}_{\text{RI}}, \theta^{(1)}_{\text{COV}\times\text{OUT}}, \theta^{(1)}_{\text{RI}\times\text{OUT}}]'$,

and $\mathbf{X}^{(1)} = hcat(1_{8\times1}, (1_{2\times1} \otimes [0\ 0\ 1\ 1]'), (1_{4\times1} \otimes [0\ 1]'), ([0\ 1]' \otimes 1_{4\times1}), (1_{2\times1} \otimes [0\ 0\ 0\ 1]'), [0\ 0\ 0\ 0\ 0\ 1\ 0\ 1]')$.

EM Algorithm. The E step is given in (3.1). The M step fits $\mathbf{Y}^*$ using iterative proportional fitting (IPF) with fixed margins COV × OUT and OUT × RI. The estimated cell probabilities (× 100) on the COV × OUT margin are $\mathbf{P} = 100(1_{1\times2} \otimes \mathbf{I}_4)\hat{\mathbf{U}}^*/(1_{1\times8}\hat{\mathbf{U}}^*) = [39.39\ 13.94\ 11.82\ 34.85]'$.

Observed Information Matrix. The observed information matrix is computed via (3.8) with $\mathbf{G}^{(1)} = \mathbf{I}_8$, $\dot{\mathbf{H}}^{(1)} = 1_{8\times1}$ and $\ddot{\mathbf{H}}^{(1)} = 0_{8\times1}$. Applying the delta method to the inverse of the observed information matrix gives standard errors for $\mathbf{P}$ of $[4.16\ 3.65\ 2.35\ 3.10]'$. Because $\mathbf{P}$ is not a function of $\hat{\phi}$, the asymptotic standard errors are correct under the multinomial distribution.

4.3 A COUPLED LOG-LINEAR/LOGIT MODEL

Fay (1986) and Baker and Laird (1988) proposed recursive systems of loglinear and logit models for partially observed categorical variables.

Random Variables, Incomplete Data, and Composite Link. These are described in Example 4.2.

Model. The overall model consists of a log-linear model for the margin of interest with variables COV and OUT (the margin model), coupled with a logit model for RI, given COV and OUT (the nonresponse model). When nonresponse is nonignorable, the overall model is $(\mathbf{U}^*) = \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)} + ((\mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)}) \cdot \mathbf{Z}^{(2)}) - \log(1 + \exp(\mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)}))$, where $\boldsymbol{\theta}^{(1)} = [\phi, \theta^{(1)}_{\text{COV}}, \theta^{(1)}_{\text{OUT}}, \theta^{(1)}_{\text{COV}\times\text{OUT}}]'$, $\boldsymbol{\theta}^{(2)} = [\theta^{(2)}_{\text{RI}}, \theta^{(2)}_{\text{RI}\times\text{OUT}}]'$, $\mathbf{X}^{(1)} = 1_{2\times1} \otimes \tilde{\mathbf{X}}^{(1)}$, $\tilde{\mathbf{X}}^{(1)} = hcat([1\ 1\ 1\ 1]', [0\ 0\ 1\ 1]', [0\ 1\ 0\ 1]', [0\ 0\ 0\ 1]')$, $\mathbf{X}^{(2)} = hcat(1_{8\times1}, (1_{4\times1} \otimes [0\ 1]'))$, and $\mathbf{Z}^{(2)} = [0\ 0\ 0\ 0\ 1\ 1\ 1\ 1]'$.

EM Algorithm. The E step is given by (3.1). Because the likelihood for the complete data factors into separate components involving $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$, on the M step, I separately estimate expected cell counts for the margin and nonresponse models. The marginal counts for the complete data are $\mathbf{Y}^*_1 = \mathbf{M}\mathbf{Y}^*$, where $\mathbf{M} = 1_{2\times1} \otimes \mathbf{I}_4$. To obtain estimated expected cell counts for the margin model, $\hat{\mathbf{U}}^{*(1)}$, I fit $\mathbf{Y}^*_1$ using IPF with fixed margin COV × OUT. To obtain estimated expected cell counts for the nonresponse model, $\hat{\mathbf{U}}^{*(2)}$, I fit $\mathbf{Y}^*$ using IPF with fixed margins COV × OUT and OUT × RI. The estimated expected counts for the combination of models are $\mathbf{U}^* = \mathbf{U}^{*(2)} \cdot (\mathbf{M}'\mathbf{U}^{*(1)})/(\mathbf{M}'\mathbf{M}\,\mathbf{U}^{*(2)})$. After convergence, I recover parameter estimates by $\hat{\boldsymbol{\theta}}^{(1)} = (\tilde{\mathbf{X}}^{(1)\prime}\tilde{\mathbf{X}}^{(1)})^{-1}\tilde{\mathbf{X}}^{(1)\prime}\log(\hat{\mathbf{U}}^{*(1)})$ and $\hat{\boldsymbol{\theta}}^{(2)} = (\tilde{\mathbf{X}}^{(2)\prime}\tilde{\mathbf{X}}^{(2)})^{-1}\tilde{\mathbf{X}}^{(2)\prime}\log(\Omega^{(2)}/(1_{8\times1} - \Omega^{(2)}))$, where $\tilde{\mathbf{X}}^{(2)} = \mathbf{X}^{(2)} \cdot (\mathbf{Z}^{(2)} \otimes 1_{1\times2})$ and $\Omega^{(2)} = \hat{\mathbf{U}}^{*(2)}/\mathbf{M}'\mathbf{M}\hat{\mathbf{U}}^{*(2)}$. The estimate of $\theta^{(1)}_{\text{COV}\times\text{OUT}}$ is 2.120.

Observed Information Matrix. The observed information matrix is computed via (3.8) with $\mathbf{G}^{(1)} = \mathbf{G}^{(2)} = \mathbf{I}_8$, $\dot{\mathbf{H}}^{(1)} = 1_{8\times1}$, $\ddot{\mathbf{H}}^{(1)} = 0_{8\times1}$, $\dot{\mathbf{H}}^{(2)} = \mathbf{Z}^{(2)} - \Pi^{(2)}$, $\ddot{\mathbf{H}}^{(2)} = \Pi^{(2)} \cdot (\Pi^{(2)} - 1)$, where $\Pi^{(2)} = \exp(\mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)})/(1 + \exp(\mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)}))$. The standard error of $\hat{\theta}^{(1)}_{\text{COV}\times\text{OUT}}$ is .3367.

4.4 A LATENT CLASS MODEL FOR THE RESULTS OF TWO DIAGNOSTIC TESTS

This example is from Hui and Walter (1980).

Random Variables. For this discussion the random variables are called TEST1 (outcome of Tine test, either positive or negative), TEST2 (outcome of Mantoux test, either positive or negative), GROUP (1 or 2), and DISEASE (presence or absence of tuberculosis according to a latent gold standard).

Complete Data. The complete data are a $2 \times 2 \times 2 \times 2$ cross-classification of TEST1, TEST2, GROUP, and DISEASE, reshaped as a 16×1 array.

Incomplete Data. The incomplete data are $\mathbf{Y} = [14\ 4\ 9\ 528\ 887\ 31\ 37\ 367]'$, which represents a $2 \times 2 \times 2$ cross-classification of TEST1, TEST2, and GROUP, reshaped as an 8×1 vector. The realizations of the latent variable DISEASE are not observed.

Composite Link. The composite link matrix is $\mathbf{C} = [1\ 1] \otimes \mathbf{1}_{8\times 1}$.

Model. The model for conditional independence of tests, given presence or absence of tuberculosis, is $\log(\mathbf{U}^*) = \log(\mathbf{N}) + \log(\mathbf{Z}^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)}) + \log(\mathbf{Z}^{(2)} + \mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)}) + \log(\mathbf{Z}^{(3)} + \mathbf{X}^{(3)}\boldsymbol{\theta}^{(3)})$, where $\boldsymbol{\theta}^{(1)}$ = [proportion with disease in group 1, proportion with disease in group 2]', $\boldsymbol{\theta}^{(2)}$ = [false negative rate of test 1, false positive rate of test 1]', $\boldsymbol{\theta}^{(3)}$ = [false negative rate of test 2, false positive rate of test 2]', $\mathbf{N} = \mathbf{1}_{2\times 1} \otimes [555\ 1322]' \otimes \mathbf{1}_{4\times 1}$, $\mathbf{Z}^{(1)} = [0\ 1]' \otimes \mathbf{1}_{8\times 1}$, $\mathbf{Z}^{(2)} = [1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 1\ 1]'$, $\mathbf{Z}^{(3)} = [1\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 0\ 1\ 0\ 1\ 0\ 1\ 0\ 1]'$, $\mathbf{X}^{(1)} = [1\ -1]' \otimes \mathbf{I}_2 \otimes \mathbf{1}_{4\times 1}$, $\mathbf{X}^{(2)} = block((\mathbf{1}_{2\times 1} \otimes [-1\ 1\ -1\ 1]'), (\mathbf{1}_{2\times 1} \otimes [1\ 1\ -1\ -1]'))$, and $\mathbf{X}^{(3)} = block((\mathbf{1}_{4\times 1} \otimes [-1\ 1]'), (\mathbf{1}_{4\times 1} \otimes [1\ -1]'))$.

EM Algorithm. The *E* step is given by (3.1). The *M* step is $\widehat{\boldsymbol{\theta}}^{(k)} = (\mathbf{Y}'[\mathbf{X}^{(k)} = 1])/(\mathbf{Y}'|\mathbf{X}^{(k)}|)$ for $k = 1, 2, 3$.

Observed Information Matrix. The observed information matrix is computed via (3.8), with $\mathbf{G}^{(k)} = \mathbf{I}_{16}$, $\dot{\mathbf{H}}^{(k)} = \mathbf{1}_{16\times 1}/(\mathbf{Z}^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)})$, and $\ddot{\mathbf{H}}^{(k)} = -\mathbf{1}_{16\times 1}/(\mathbf{Z}^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)})^2$, for $k = 1, 2$. Because the model fits perfectly, the observed information matrix equals the expected information matrix given in Hui and Walter (1980).

4.5 LINEAR MODELS FOR MISCLASSIFIED DATA WITH A VALIDATION SAMPLE

This example was discussed in Chen (1979) and Palmgren and Ekholm (1987).

Random Variables. For this discussion the random variables are called JOINT-P (joint pain as determined by a physician's exam, either yes or no), JOINT-I (joint pain as determined by interview, either yes or no), SEX (male or female), and STRATUM (severity of arthritis or rheumatism, 1, 2, or 3). The true probability of joint pain is ascertained by JOINT-P; a proxy for joint pain is JOINT-I.

Complete Data. The complete data are a $2 \times 2 \times 2 \times 3 \times 2$ cross-classification of JOINT-P, JOINT-I, SEX, STRATUM, and sample, reshaped as 48×1 vector.

Incomplete Data. The incomplete data for one random sample are $\mathbf{Y}_{RS1}$ = [65 2 5 2 24 5 12 10 35 16 20 41 64 3 4 1 23 2 11 5 36 7 25 60]', which is a $2 \times 2 \times 3 \times 2$ cross-classification of JOINT-P, JOINT-I, SEX, and STRATUM, reshaped as a 24×1 vector. The incomplete data for the other random sample are $\mathbf{Y}_{RS2}$ = [69 10 25 23 45 70 69 8 27 20 35 75]', which is a $2 \times 2 \times 3$ cross-classification of

JOINT-I, SEX, and STRATUM, reshaped as a 16×1 vector. Thus, all of the incomplete data are given by $\mathbf{Y} = vcat(\mathbf{Y}_{RS1}, \mathbf{Y}_{RS2})$.

Composite Link. The composite link matrix is $\mathbf{C} = block(\mathbf{I}_{24}, (\mathbf{I}_{12} \otimes \mathbf{1}_{1\times 2}))$.

Model. Palmgren and Ekholm (1987) fit a model in which the true probability of joint pain (JOINT-P), the false negative rate, and the false positive rate depend on severity of arthritis and rheumatism but not sex. In this notation the model is $\log(\mathbf{U}^*) = \log(\mathbf{N}) + \log(\mathbf{Z}^{(1)} + \mathbf{X}^{(1)}\boldsymbol{\theta}^{(1)}) + \log(\mathbf{Z}^{(2)} + \mathbf{X}^{(2)}\boldsymbol{\theta}^{(2)})$, where $\boldsymbol{\theta}^{(1)}$ = [true probability of joint pain in stratum 1, true probability of joint pain in stratum 2, true probability of joint pain in stratum 3]′, $\boldsymbol{\theta}^{(2)}$ = [false negative rate in stratum 1, false negative rate in stratum 2, false negative rate in stratum 3, false positive rate in stratum 1, false positive rate in stratum 2, false positive rate in stratum 3]′, $\mathbf{N} = [74\ 51\ 112\ 72\ 41\ 128\ 79\ 48\ 115\ 77\ 47\ 112]' \otimes \mathbf{1}_{4\times 1}$, $\mathbf{Z}^{(1)} = \mathbf{1}_{24\times 1} \otimes [0\ 1]'$, $\mathbf{Z}^{(2)} = \mathbf{1}_{12\times 1} \otimes [1\ 0\ 0\ 1]'$, $\mathbf{X}^{(1)} = \mathbf{1}_{4\times 1} \otimes \mathbf{I}_3 \otimes [1\ -1\ 1\ -1]'$, and $\mathbf{X}^{(2)} = hcat((\mathbf{1}_{4\times 1} \otimes \mathbf{I}_3 \otimes [-1\ 0\ 1\ 0]'), (\mathbf{1}_{4\times 1} \otimes \mathbf{I}_3 \otimes [0\ 1\ 0\ -1]'))$. The M step is $\widehat{\boldsymbol{\theta}}^{(1)} = (\mathbf{Y}'[\mathbf{X}^{(1)} = 1])/(\mathbf{Y}'|\mathbf{X}^{(1)}|)$ and $\widehat{\boldsymbol{\theta}}^{(2)} = (\mathbf{Y}'[\mathbf{X}^{(2)} = 1])/(\mathbf{Y}'|\mathbf{X}^{(2)}|)$.

Observed Information Matrix. The observed information matrix is computed via (3.8), with $\mathbf{G}^{(k)} = \mathbf{I}_{48}$, $\dot{\mathbf{H}}^{(k)} = \mathbf{1}_{48\times 1}/(\mathbf{Z}^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)})$, and $\ddot{\mathbf{H}}^{(k)} = -\mathbf{1}_{48\times 1}/(\mathbf{Z}^{(k)} + \mathbf{X}^{(k)}\boldsymbol{\theta}^{(k)})^2$, for $k = 1, 2$. Standard errors based on the expected information matrix are identical to those obtained by Palmgren and Ekholm (1987). Standard errors based on the observed information matrix differ in the second or third significant digit.

ACKNOWLEDGMENTS

The author thanks Charles Brown, Nan Laird, Thomas Louis, Blossom Patterson, and Philip Prorok for helpful suggestions.

REFERENCES

Baker, S. G. (1992), "Simplifying Variance Computation by Substituting a Poisson for a Multinomial Distribution," unpublished manuscript, submitted to *The American Statistician*.

Baker, S. G., and Chu, K. (1990), "Evaluating Screening for the Early Detection and Treatment of Cancer Without Using a Randomized Control Group," *Journal of the American Statistical Association*, 85, 321–327.

Baker, S. G., and Laird, N. M. (1988), "Regression Analysis for Categorical Variables With Outcome Subject to Nonignorable Nonresponse," *Journal of the American Statistical Association*, 83, 62–69.

Baker, S. G., Wax, Y., and Patterson, B. H. (1992), "Regression Analysis of Grouped Survival Data in the Presence of Informative Censoring and Double Sampling," unpublished manuscript, submitted to *Biometrics*.

Burn, R. (1983), "Fitting a Logit Model to Data With Classification Errors," *GLIM Newsletter*, 8, 44–47.

Chen, T. T. (1979), "Loglinear Models for Categorical Data With Misclassification and Double Sampling," *Journal of the American Statistical Association*, 74, 481–488.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–38.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

Efron, B., and Hinkley, D. V. (1977), "Assessing the Accuracy of the Maximum Likelihood Estimator: Observed Versus Expected Fisher Information," *Biometrika*, 65, 457–487.

Espeland, M. A. (1986), "A General Class of Models for Discrete Multivariate Data," *Communications in Statistics: Simulation and Computation*, 15, 405–424.

Espeland, M. A., and Odoroff, C. L. (1985), "Log-Linear Models for Doubly Sampled Categorical Data Fitted by the EM Algorithm," *Journal of the American Statistical Association*, 80, 663–670.

Fay, R. E. (1986), "Causal Models for Patterns of Nonresponse," *Journal of the American Statistical Association*, 81, 354–365.

Green, P. J. (1984), "Iteratively Reweighted Least Squares for Maximum Likelihood Estimation, and Some Robust and Resistant Alternatives," *Journal of the Royal Statistical Society*, Ser. B, 46, 149–192.

Hartley, H. O. (1958), "Maximum Likelihood Estimation From Incomplete Data," *Biometrics*, 14, 174–194.

Hartley, H. O., and Hocking, R. R. (1971), "The Analysis of Incomplete Data," *Biometrics*, 27, 783–823.

Hui, S. L., and Walter, S. D. (1980), "Estimating the Error Rates of Diagnostic Tests," *Biometrics*, 36, 167–171.

Ibrahim, J. G. (1990), "Incomplete Data in Generalized Linear Models," *Journal of the American Statistical Association*, 85, 765–769.

Jorgensen, B. (1983), "Maximum Likelihood Estimation and Large-Sample Inference for Generalized Linear and Nonlinear Regression Models," *Biometrika*, 70, 19–28.

Jorgensen, M. (1987), "Jackknifing Fixed Points of Iteration," *Biometrika*, 74, 207–211.

Little, R. J. A., and Rubin, D. B. (1987), *Statistical Analysis With Missing Data*, New York: John Wiley.

Louis, T. A. (1982), "Finding the Observed Information Matrix When Using the EM Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 44, 226–233.

Meilijson, I. (1989), "A Fast Improvement of the EM Algorithm on Its Own Terms," *Journal of the Royal Statistical Society*, Ser. B, 51, 127–138.

Meng, X., and Rubin, D. B. (1991), "Using EM to Obtain Asymptotic Variance–Covariance Matrices: The SEM Algorithm," *Journal of the American Statistical Association*, 86, 899–909.

Palmgren, J. (1981), "The Fisher Information Matrix for Log Linear Models Arguing Conditionally on Observed Explanatory Variables," *Biometrika*, 68, 563–566.

Palmgren, J., and Ekholm, A. (1987), "Exponential Family Non-linear Models for Categorical Data With Errors of Observation," *Applied Stochastic Modeling and Data Analysis*, 3, 111–124.

Rao, C. R. (1965), *Linear Statistical Inference and Its Applications*, New York: John Wiley.

Redner, R. A., and Walker, H. F. (1984), "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195–239.

Tanner, M. A., and Wong, W. H. (1987), "The Calculation of Posterior Distributions by Data Augmentation" (With Discussion), *Journal of the American Statistical Association*, 82, 528–550.

Thompson, R., and Baker, R. J. (1981), "Composite Link Functions in Generalized Linear Models," *Applied Statistics*, 30, 125–131.

Wei, G. C. C., and Tanner, M. A. (1990), "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms," *Journal of the American Statistical Association*, 85, 699–704.