Finding the Observed Information Matrix when Using the EM Algorithm
Author(s): Thomas A. Louis
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 44, No. 2 (1982), pp. 226-233
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2345828
Accessed: 24-06-2019 08:03 UTC

J. R. Statist. Soc. B (1982),
44, *No.* 2, *pp.* 226–233

Finding the Observed Information Matrix when Using the *EM* Algorithm

By THOMAS A. LOUIS

Harvard School of Public Health, Mass., USA

[Received July 1980. Revised January 1981]

SUMMARY

A procedure is derived for extracting the observed information matrix when the *EM* algorithm is used to find maximum likelihood estimates in incomplete data problems. The technique requires computation of a complete-data gradient vector or second derivative matrix, but not those associated with the incomplete data likelihood. In addition, a method useful in speeding up the convergence of the *EM* algorithm is developed. Two examples are presented.

Keywords: EM ALGORITHM; OBSERVED INFORMATION; MAXIMUM LIKELIHOOD; SPEEDING CONVERGENCE

1. INTRODUCTION

THE *EM* algorithm for finding maximum likelihood estimates (MLE's) is a powerful numerical technique useful in contexts ranging from standard incomplete data problems (e.g. censored and truncated), to iteratively reweighted least squares analysis and empirical Bayes models. Several analyses of its properties with examples and cautions for its use have appeared in the literature (see Louis *et al.*, 1976; Sundberg, 1976; Dempster *et al.*, 1977; Laird, 1978, and the references thereof). These discussions will not be repeated here.

The primary conceptual power of this iterative algorithm lies in converting a maximization problem involving a complicated likelihood, into a sequence of "pseudo-complete" problems, where at each step the updated parameter estimates can be obtained in a closed form (or at least in a straightforward manner). Unlike Newton–Raphson or Fletcher–Powell techniques, no gradients or curvature matrices need to be derived. Unfortunately this conceptual and analytic simplification does not appear to provide a means of estimating the information matrix associated with the maximum likelihood estimates. There have been, however, solutions published for a few special cases (see Hartley and Hocking, 1971, for example). Of course, an alternative to the method contained herein is use of a derivative-free function-maximizing algorithm.

In this paper, a technique for computing the observed information (see Efron and Hinkley 1978) within the *EM* framework will be presented. It requires computation of the complete-data gradient and second derivative matrix and can be imbedded quite simply in the *EM* iterations. In addition, a technique potentially useful for improving the speed of convergence of the algorithm is developed. Background details from the cited references will not be reproduced, but some basic examples are included. More complicated applications (where the *EM* is most useful) follow directly from the theory, but quickly become notationally opaque. For the type of application which generated the current research, see Turnbull (1974), Louis *et al.* (1978), Dinse (1982) and Laird and Louis (1982).

2. THE ML PROBLEM AND *EM* ALGORITHM

The objective is to find the MLE for a p-dimensional parameter θ in a space Θ. The underlying probability model induces a density or mass function $f(x \mid \theta)$ on a sample space χ, where $x = (x_1, ..., x_n)^T$. Instead of observing $x \in \chi$, one observes the value of a measurable function $Y(x) = y \in Y$. The MLE $(\hat{\theta})$ is to be found using the data y.

 0035–9246/82/44226$2.00

The *EM* method is only attractive in situations where finding the complete data MLE and either the observed or the expected information matrix would be straightforward, but the problem based on the incomplete data (Y) requires an iterative solution. Typically the complete data are from an exponential family. The algorithm operates as follows. Let

$$\lambda(x, \theta) = \log\{f(x \mid \theta)\},$$

$$\lambda^*(y, \theta) = \log\{f_Y(y \mid \theta)\} = \log\left\{\int_R f_X(x \mid \theta)\, d\mu(x)\right\},$$

where $R = \{x: y(x) = y\}$, and $\mu(x)$ is a dominating measure. The case when the dimension of R is less than n requires special notation, which will not be developed here.

Now, instead of maximizing λ^* directly, the *EM* algorithm proceeds by using an initial estimate $\theta^{(0)}$ and solving the pseudo-complete data problem:

$$\underset{\theta \in \Theta}{\text{maximize}}\, E_{\theta^{(0)}}[\lambda(X, \theta) \mid X \in R].$$

The maximizing value for this pseudo-complete data problem is called $\theta^{(1)}$ and the iteration is continued until $\| \theta^{(v+1)} - \theta^{(v)} \|$ is sufficiently small or some other convergence criterion is satisfied. Under conditions specified in Sundberg (1976), $\theta^{(v)} \to \hat{\theta}$, the MLE. Notice that the *EM* algorithm induces the map $\theta^{(v+1)} = g(\theta^{(v)})$, which will be used in Section 5.

3. Estimating the Information

3.1. *The General Case*

We assume that the regularity conditions in Zacks (1971, Chapter 5) hold. These guarantee that the MLE solves the gradient equation (3.1) and that the Fisher information exists. To see how to compute the observed information in the *EM*, let $S(x, \theta)$ and $S^*(y, \theta)$ be the gradient vectors of λ and λ^* respectively and $B(x, \theta)$ and $B^*(y, \theta)$ be the negatives of the associated second derivative matrices. Then by straightforward differentiation (see the Appendix):

$$\begin{aligned} &S^*(y, \theta) = E_\theta[S(X, \theta) \mid X \in R], \\ &S^*(y, \hat{\theta}) = 0, \end{aligned} \tag{3.1}$$

$$I_Y(\theta) = E_\theta\{B(X, \theta) \mid X \in R\} - E_\theta\{S(X, \theta)\, S^{\mathrm{T}}(X, \theta) \mid X \in R\} + S^*(y, \theta)\, S^{*\mathrm{T}}(y, \theta). \tag{3.2}$$

The first term in (3.2) is the conditional expected full data observed information matrix, while the last two produce the expected information for the conditional distribution of X given $X \in R$. That is, using a simplified notation:

$$I_Y = I(\hat{\theta}) = I_X - I_{X \mid Y}, \tag{3.3}$$

which is an application of the missing information principle (Woodbury, 1977) to the observed information. Notice that all of these conditional expectations can be computed in the *EM* algorithm using only S and B, which are the gradient and curvature for a complete-data problem. Of course, they need be evaluated only on the last iteration of the *EM* procedure, where S^* is zero.

Efron and Hinkley (1978) define I_Y as the observed information and show that in most cases it is a more appropriate measure of information than the *a priori* expectation $E_\theta[B^*(Y, \theta)]$. It is certainly easier to compute.

3.2. *The Independence Case*

When $X_1, \ldots, X_n$ are independent but not necessarily identically distributed and $Y_i(X) = Y_i(X_i)$, the λ, λ^*, S, S^* and B, B^* become summations and $R = R_1 \times R_2 \times \ldots \times R_n$.

Thus

$$S^*(y,\theta) = \sum_{i=1}^{n} S_i^*(y_i,\theta)$$

$$= \sum_{i=1}^{n} E_\theta\{S_i(X_i,\theta) \mid X_i \in R_i\} \tag{3.1'}$$

and

$$I_Y = \sum_{i=1}^{n} E_{\hat\theta}\{B_i(X_i,\hat\theta) \mid X_i \in R_i\}$$

$$- \sum_{i=1}^{n} E_{\hat\theta}\{S_i(X_i,\hat\theta)\, S_i^{\mathrm{T}}(X_i,\hat\theta) \mid X \in R_i\}$$

$$-2\sum_{i<j}^{n} E_{\hat\theta}\{S_i(x_i,\hat\theta) \mid X_i \in R_i\}\, E_{\hat\theta}\{S_j(X_j,\hat\theta) \mid X_j \in R_j\}^{\mathrm{T}}. \tag{3.2'}$$

When each X_i is an indicator vector for multinomial distribution, I_Y reduces to

$$\sum_{i=1}^{n} S(\hat X_i,\hat\theta)\, S^{\mathrm{T}}(\hat X_i,\hat\theta), \tag{3.4}$$

where

$$\hat X_i = E_{\hat\theta}[X_i \mid X_i \in R_i].$$

4. EXAMPLES

4.1. *Example from Dempster* et al. (1977)

Here, θ is to be estimated from the multinomial distribution:

$$\{(\tfrac{1}{2}+\tfrac{1}{4}\theta), \tfrac{1}{4}(1-\theta), \tfrac{1}{4}(1-\theta), \tfrac{1}{4}\theta\}, \quad 0 \leqslant \theta \leqslant 1.$$

With Y_1, Y_2, Y_3, Y_4 as the frequencies, let

$$Y_1 = X_1 + X_2, \quad Y_2 = X_3, \quad Y_3 = X_4, \quad Y_4 = X_5,$$

where X is multinomial with parameters

$$\{\tfrac{1}{2}, \tfrac{1}{4}\theta, \tfrac{1}{4}(1-\theta), \tfrac{1}{4}(1-\theta), \tfrac{1}{4}\theta\}.$$

Therefore, if X were observed,

$$\hat\theta = \frac{X_2 + X_5}{(X_2 + X_3 + X_4 + X_5)}.$$

The MLE can be found by solving a quadratic and with data $Y = (125, 18, 20, 34)$, $\hat\theta = 0{\cdot}6268215\ldots$.

Alternatively, the *EM* algorithm can be used where

$$X_2^{(v+1)} = \frac{\frac{1}{4}\theta^{(v)}}{\frac{1}{2}+\frac{1}{4}\theta^{(v)}}, \quad Y_1 = \frac{\theta^{(v)}}{2+\theta^{(v)}}\, Y_1, \quad X_1^{(v+1)} = Y_1 - X_2^{(v+1)},$$

and $X_k = Y_{k-1}$, $k = 3, 4, 5$. Here

$$\theta^{(v+1)} = \frac{Y_1\,\theta^{(v)} + (2+\theta^{(v)})\, Y_4}{Y_1\,\theta^{(v)} + (2+\theta^{(v)})(Y_2 + Y_3 + Y_4)}$$

$$= \frac{X_2^{(v+1)} + X_5}{X_2^{(v+1)} + X_3 + X_4 + X_5}.$$

Starting at $\theta^{(0)} = 0{\cdot}60$, the first few iterates are ($d^{(v)} = \theta^{(v)} - \theta^{(v-1)}$).

$$\theta^{(0)} = 0{\cdot}600000, \qquad d^{(1)} = 0{\cdot}02318800,$$
$$\theta^{(1)} = 0{\cdot}623188, \qquad d^{(2)} = 0{\cdot}00314975,$$
$$\theta^{(2)} = 0{\cdot}626338, \qquad d^{(3)} = 0{\cdot}00041927,$$
$$\theta^{(3)} = 0{\cdot}626757,$$
$$\theta^{(4)} = 0{\cdot}626812,$$
$$\theta^{(\infty)} = 0{\cdot}6268215\ldots.$$

Here

$$S(X, \theta) = \frac{X_2 + X_5}{\theta} - \frac{X_3 + X_4}{1-\theta},$$

$$B(X, \theta) = \frac{X_2 + X_5}{\theta^2} + \frac{X_3 + X_4}{(1-\theta)^2}.$$

Direct computation is especially straightforward, since (3.4) applies. Nevertheless, applying (3.2′) gives a measure of the loss of information, with

$$I_Y = 435{\cdot}3 - 57{\cdot}8 = 377{\cdot}5.$$

This information is associated with an *effective* sample size of $88{\cdot}3$ [$= 377{\cdot}5\hat{\theta}(1-\hat{\theta})$].

4.2. *Example: Mixture of Two Normals*

The case of equal, known variances will be considered. Data $Y_1, \ldots, Y_n$ are observed and are known to be i.i.d. from a mixture of two normal distributions. Assuming $\sigma^2 = 1$, the density of each co-ordinate of Y is

$$f_Y(y \mid \mu_0, \mu_1, \varepsilon) = (1-\varepsilon)\,\phi(y-\mu_0) + \varepsilon\phi(y-\mu_1), \quad -\infty < \mu_0, \mu_1 < \infty, \quad 0 \leqslant \varepsilon \leqslant \tfrac{1}{2},$$

where

$$\phi(z) = \frac{1}{\sqrt{(2\pi)}} e^{-z^2/2}.$$

This can be considered a missing data problem by letting $X = (Y, Z)$, where

$$Z = \begin{cases} 0 \\ 1 \end{cases} \text{according as } Y \text{ is from } \begin{matrix} N(\mu_0, 1) \\ N(\mu_1, 1) \end{matrix}.$$

The vector $Z = (Z_1, \ldots, Z_n)$ is unobserved. Implementation of the *EM* algorithm is straightforward. Let

$$w_i = w(Y_i, \mu_0, \mu_1, \varepsilon) = E[Z_i \mid Y_i, \mu_0, \mu_1, \varepsilon)]$$
$$= \frac{\varepsilon\phi(Y_i - \mu_1)}{\varepsilon\phi(Y_i - \mu_1) + (1-\varepsilon)\,\phi(Y_i - \mu_0)}$$

and

$$w_i^{(v)} = w(Y_i, \mu_0^{(v)}, \mu_1^{(v)}, \varepsilon^{(v)}).$$

Then, the parameters are updated by

$$\varepsilon^{(v+1)} = \frac{1}{n}\sum_{i=1}^{n} w_i^{(v)},$$

$$\mu_0^{(v+1)} = \{n(1-\varepsilon^{(v+1)})\}^{-1}\sum_{i=1}^{n}(1-w_i^{(v)})\,Y_i,$$

$$\mu_1^{(v+1)} = \{n\varepsilon^{(v+1)}\}^{-1}\sum_{i=1}^{n} w_i^{(v)}\,Y_i.$$

The complete data gradient $S(X,\theta)$ is needed to compute the observed information. This likelihood models a two-stage experiment, where first a population is picked by a Bernoulli experiment, and then a normal variate is generated. Therefore,

$$\lambda(X;\,\mu_0,\mu_1,\varepsilon) = Z\log(\varepsilon)+(1-Z)\log(1-\varepsilon) + (1-Z)\log(\phi(Y_i-\mu_0))+Z\log(\phi(Y-\mu_1))$$

so $S(X_i;\,\mu_0,\mu_1,\varepsilon)$ is

$$\frac{\partial}{\partial\mu_0} = (1-Z_i)(Y_i-\mu_0),$$

$$\frac{\partial}{\partial\mu_1} = Z_i(Y_i-\mu_i),$$

$$\frac{\partial}{\partial\varepsilon} = \frac{Z_i}{\varepsilon}-\frac{1-Z_i}{1-\varepsilon},$$

$$B(X_i,\mu_0,\mu_1,\varepsilon) = \begin{pmatrix} 1-Z_i & 0 & 0 \\ 0 & Z_i & 0 \\ 0 & 0 & (Z_i/\varepsilon^2)+(1-Z_i)/(1-\varepsilon)^2 \end{pmatrix}$$

and

$$S^*(Y_i,\mu_0,\mu_1,\varepsilon) = \begin{pmatrix} (1-w_i)(Y_i-\mu_0) \\ w_i(Y_i-\mu_1) \\ \dfrac{w_i}{\varepsilon}-\dfrac{(1-w_i)}{1-\varepsilon} \end{pmatrix}.$$

To see how the procedure works, 500 observations were generated using $\mu_0 = 2$, $\mu_1 = 0$ and $\varepsilon = 0.5$†. From these data:

$$\hat{\mu}_0 = 1.970, \quad \hat{\mu}_1 = -0{\cdot}042, \quad \hat{\varepsilon} = 0{\cdot}483,$$

$$I_X = \begin{pmatrix} 241{\cdot}7 & 0 & 0 \\ 0 & 258{\cdot}3 & 0 \\ 0 & 0 & 2002{\cdot}2 \end{pmatrix}$$

and I_Y is symmetric with lower triangle:

$$\begin{pmatrix} 136{\cdot}9 & & \\ -20{\cdot}6 & 159{\cdot}7 & \\ 218{\cdot}5 & 219{\cdot}1 & 1131{\cdot}2 \end{pmatrix}.$$

I_Y can be inverted to find the covariance matrix of ($\hat{\mu}_0$, $\hat{\mu}_1$ and $\hat{\varepsilon}$).

† I have chosen a case where convergence is rapid and the chance of converging to a non-global maximum is small.

5. SPEEDING UP CONVERGENCE OF THE *EM*

5.1. *The Basic Approximation*

Following Louis *et al.* (1976) and Sundberg (1976), if $J(\theta)$ is the Jacobian of the map $\theta^{(v+1)} = g(\theta^{(v)})$, then local to the MLE $\hat{\theta}(= \theta^{(\infty)})$, with $d^{(v)} = \theta^{(v)} - \theta^{(v-1)}$

$$d^{(v+1)} \doteq J(\theta^{(v-1)})\, d^{(v)} \doteq \left\{ \prod_{l=1}^{v} J(\theta^{(l-1)}) \right\} d^{(1)}, \quad v\text{P1}. \tag{5.1}$$

Equation (5.1) is derived by expanding $g(\theta^{(v)})$ in a first-order Taylor series about $\theta^{(v-1)}$.

From (5.1), for large j and all $v \geqslant 1$ the approximation:

$$d^{(v+j+1)} \doteq J^{v}(\hat{\theta})\, d^{(j+1)}$$

holds, and implies that, with J denoting $J(\hat{\theta})$:

$$\begin{aligned} \theta^{(\infty)} = \hat{\theta} \equiv \theta^{(j)} + \sum_{l=1}^{\infty} d^{(l+j)} \\ \doteq \theta^{(j)} + \left(\sum_{l=0}^{\infty} J^{l} \right) d^{(j+1)} \\ = \theta^{(j)} + (1-J)^{-1}\, d^{(j+1)}, \end{aligned} \tag{5.2}$$

where 1 is the $p \times p$ identity matrix and θ is p-dimensional. The last step is justified by the fact that at the MLE the eigenvalues of J are all less than one in absolute value. In fact the largest modulus of them determines the speed of convergence of the *EM* algorithm in a neighbourhood of $\hat{\theta}$. Formula (5.2) is an example of Aitken's acceleration method.

Using (5.2) there is the possibility of producing the effect of an infinite number of iterations by the following algorithm:

1. From $\theta^{(j)}$ produce $\theta^{(j+1)}$ using *EM*.
2. Estimate $J(\theta^{(j)})$ by $\hat{J}$ (see Section 5.2).
3. Compute
$$\theta_{*}^{(j+1)} = \theta^{(j)} + (1-\hat{J})^{-1}\, d^{(j+1)}.$$
4. Use $\theta_{*}^{(j+1)}$ in step 1.

5.2. *Estimating* $(1-J)^{-1}$

Using (3.3) and the results of either Louis *et al.* (1976) or Sundberg (1976), $\hat{J}$ satisfies

$$\hat{J} = (I_X - I_Y)\, I_X^{-1} = (1 - I_Y\, I_X^{-1}) \tag{5.3}$$

and so

$$(1-\hat{J})^{-1} = I_X\, I_Y^{-1}.$$

It is important to stress that the expected information for X should not be used, for there is no guarantee that subtracting I_Y from it results in a non-negative definite matrix.

The decision to use the Aitken projection should be based on the cost of inverting I_Y relative to running through a single iteration of the *EM*. Also, the approximation is useful only local to the MLE, and should not be used until some iterations have been performed. In addition, instead of inverting $(1-\hat{J})$, a finite series approximation to it $(\Sigma_{v=0}^{K} \hat{J}^{v})$ can be used in step 3. Of course, if $K = 0$, we are back to performing the *EM*. Since $\hat{J}$ is relatively inexpensive to compute, this option with moderate K is an attractive approach.

To see how the projection applies, consider Example 1 in Section 4. From the third iteration:

$$(1-\hat{J})^{-1} = I_X\, I_Y^{-1} = \frac{434{\cdot}79}{376{\cdot}95} = 1{\cdot}153442$$

and $\theta_{*}^{(3)} = \theta^{(3)} + 1{\cdot}153442 d^{(3)} = 0{\cdot}6268216$ which is closer to $\theta^{(\infty)}$ than is $\theta^{(4)}$.

In this example the *expected* X-information (420·87)—is 3·2 per cent smaller than I_X—and the projected θ value obtained using it is further from the MLE than $\theta_*^{(3)}$. The discrepancy is due to automatic conditioning on ancillary statistics by the observed, but not expected information. In the full-data problem, X_1 is ancillary, with an expected value (for $X_1 + \ldots + X_5 = 197$) of $197/2 = 98{\cdot}5$, irrespective of θ. Thus, while the expected number of observations giving information to θ is 98·5, there were (at the fourth iteration) 101·6 $(= 197 - X_1^{(3)})$ pseudo-observations doing so. The larger pseudo-sample size is more appropriate for this set of data than the *a priori* expected value of 98·5. It should be stressed, however, that the *effective* sample size is only 88·3. For amplification of these issues see Efron and Hinkley (1978).

In the second example,

$$\hat{J} = \begin{pmatrix} 0{\cdot}4336 & 0{\cdot}0798 & -0{\cdot}1091 \\ 0{\cdot}0852 & 0{\cdot}3817 & -0{\cdot}1094 \\ -0{\cdot}9040 & -0{\cdot}8482 & 0{\cdot}4350 \end{pmatrix}$$

which can be used in the projection approach. The eigenvalues of $\hat{J}$ are

$$(0{\cdot}9018, \quad 0{\cdot}3248, \quad 0{\cdot}0237),$$

and the largest of these determines the geometric rate of convergence of the *EM* algorithm (see (5.1), Louis *et al.*, 1976; Sundberg, 1976 and Dempster *et al.*, 1977).

APPENDIX

Derivation of (3.1) *and* (3.2)

We have

$$\lambda^*(y, \theta) = \log\{f_Y(y \mid \theta)\} = \log\left\{\int_R f_X(x \mid \theta) d\mu(x)\right\}.$$

Therefore, with $\lambda^{*\prime}$ indicating the gradient

$$S^*(y, \theta) = \lambda^{*\prime}(y, \theta) = \int_R f'(x \mid \theta)\, d\mu(x) \Big/ \int_R f(x \mid \theta)\, d\mu(x),$$

and by multiplying and dividing the integrand in the numerator by $f_X(x \mid \theta)$, we have (3.1). For (3.2) take an additional derivative to obtain the matrix:

$$\lambda^{*\prime\prime}(y, \theta) = \left\{\int_R f_X''(x \mid \theta)\, d\mu(x) \Big/ \int_R f_X(x \mid \theta)\, d\mu(x)\right\} - S^*(y, \theta)\, S^{*\mathrm{T}}(y, \theta).$$

By multiplying and dividing by $f_X(x \mid \theta)$ as before we obtain

$$\lambda^{*\prime\prime}(y, \theta) = E_\theta[f_X''(X \mid \theta)/f_X(X \mid \theta) \mid X \in R] - S^*(y, \theta)\, S^{*\mathrm{T}}(y, \theta),$$

which can be written

$$E_\theta[\lambda''(X, \theta) \mid X \in R] + E_\theta[S(X, \theta)\, S^{\mathrm{T}}(X, \theta) \mid X \in R] - S^*(y, \theta)\, S^{*\mathrm{T}}(y, \theta).$$

The negative of this expression is (3.2).

REFERENCES

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood estimation from incomplete data via the EM algorithm (with Discussion). *J. R. Statist. Soc.* B, **39**, 1–38.

DINSE, G. E. (1982). Non-parametric estimation for partially-complete time and type of failure data. *Biometrics*, **38**, (To appear.)

EFRON, B. and HINKLEY, D. V. (1978). The observed versus expected information. *Biometrika*, **65**, 457–487.

FLETCHER, R. and Powell, M. (1963). A rapidly convergent descent method for minimization, *Comp. J.*, **6**, 163–168.

HARTLEY, H. O. and HOCKING, R. R. (1971). The analysis of incomplete data. *Biometrics*, **14**, 174–194.

LAIRD, N. M. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Amer. Statist. Ass.*, **73**, 805–811.

LAIRD, N. M. and LOUIS, T. A. (1982). Approximate posterior distributions for incomplete data problems. *J. R. Statist. Soc.* B, **44**, 190–200.

LOUIS, T. A., ALBERT, A. and HEGHINIAN, S. (1978). Screening for the early detection of cancer—III. Estimation of disease natural history. *Math. Biosci.*, **40**, 111–144.

LOUIS, T. A., HEGHINIAN, S., ALBERT, A. (1976). Maximum likelihood estimation using pseudo-data interactions. *Boston University Research Report No. 2–76.*

SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, **1**, 49–58.

——— (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Comm. Statist.*, **B5**(1), 55–64.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Ass.*, **69**, 169–173.

WOODBURY, M. A. (1971). Discussion of paper by Hartley and Hocking. *Biometrics*, **27**, 808–817.

ZACKS, S. (1971). *The Theory of Statistical Inference*, pp. 230 *et seq.* New York: Wiley.